# HOPC: Histogram of Oriented Principal Components of 3D Pointclouds for Action Recognition

Hossein Rahmani, Arif Mahmood, Du Q Huynh, and Ajmal Mian

Computer Science and Software Engineering, The University of Western Australia,
35 Stirling Highway, Crawley, WA 6009 Australia

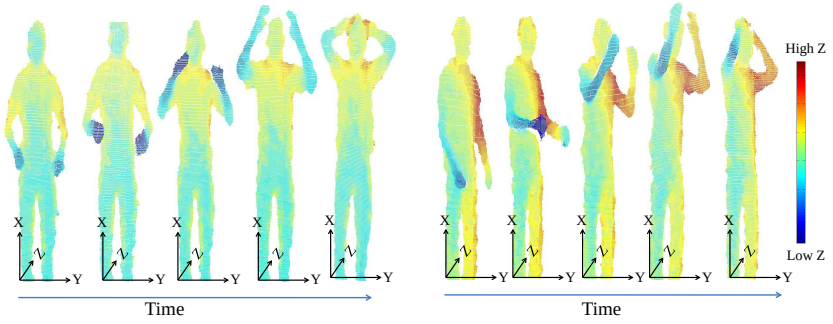
**Abstract.** Existing techniques for 3D action recognition are sensitive to viewpoint variations because they extract features from depth images which change significantly with viewpoint. In contrast, we directly process the pointclouds and propose a new technique for action recognition which is more robust to noise, action speed and viewpoint variations. Our technique consists of a novel descriptor and keypoint detection algorithm. The proposed descriptor is extracted at a point by encoding the Histogram of Oriented Principal Components (HOPC) within an adaptive spatio-temporal support volume around that point. Based on this descriptor, we present a novel method to detect Spatio-Temporal Key-Points (STKPs) in 3D pointcloud sequences. Experimental results show that the proposed descriptor and STKP detector outperform state-of-the-art algorithms on three benchmark human activity datasets. We also introduce a new multiview public dataset and show the robustness of our proposed method to viewpoint variations.

**Keywords:** Spatio-temporal keypoints, multiview action dataset

## 1 Introduction

Human action recognition has many applications in smart surveillance, human-computer interaction and sports. The Kinect and other depth cameras have become popular for this task because depth sequences do not suffer from the problems induced by variations in illumination and clothing texture. However, the presence of occlusion, sensor noise and most importantly viewpoint variations still make action recognition a challenging task.

Designing an efficient depth sequence representation is an important task in many computer vision problems. Most existing action recognition techniques (e.g., [4, 21, 38]) treat depth sequences the same way as color videos and use color-based action recognition methods. However, while these methods are suitable for color video sequences, simply extending them to depth sequences may not be optimal [19]. Information captured by depth cameras actually allows geometric features to be extracted to form rich descriptors. For instance, Tang et al. [27] used histograms of the normal vectors for object recognition in depth images. Given a depth image, they computed spatial derivatives, transformed them to the polar coordinates and used the 2D histograms as object descriptors. Recently, Oreifej and Liu [19] extended the same technique to the temporal dimension by adding time derivative. A downside of treating depth sequences this way is that the noise in the depth images is enhanced by the differential operations [31]. Histogramming, on the other hand, is analogous to integration and is more




(a) Front view (0º horizontal rotation) (b) Side view at 50º horizontal rotation

**Fig. 1.** Two sequences of 3D pointclouds of a subject performing the *holding head* action. Notice how the depth values (colours) have significantly changed with the change in viewpoint. Simple normalization cannot compensate for such depth variations. Existing depth based action recognition algorithms will not be accurate in such cases

resilient to the effect of noise. Furthermore, viewpoint variations are unavoidable in real scenarios. However, none of the existing 3D sensor based techniques is designed for cross-view action recognition where training is performed on sequences acquired from one view and testing is performed on sequences acquired from a significantly different view ($> 25°$).

We directly process the 3D pointcloud sequences (Fig. 1) and extract point descriptors which are robust to noise and viewpoint variations. We propose a novel descriptor, the *Histogram of Oriented Principal Components* (HOPC), to capture the local geometric characteristics around each point within a sequence of 3D pointclouds. To extract HOPC at a point $\mathbf{p}$, PCA is performed on an adaptive spatio-temporal support volume around $\mathbf{p}$ (see Fig. 2) which gives us a $3 \times 3$ matrix of eigenvectors and the corresponding eigenvalues. Each eigenvector is projected onto $m$ directions corresponding to the vertices of a *regular m-sided polyhedron* and scaled by its eigenvalue. HOPC is formed by concatenating the projected eigenvectors in decreasing order of their eigenvalues.

HOPC is used in a holistic and local setting. In the former approach, the sequence of 3D pointclouds is divided into spatio-temporal cells and HOPC descriptors of all points within a cell are accumulated and normalized to form a single cell descriptor. All cell descriptors are concatenated to form a holistic HOPC descriptor. In the latter approach, local HOPC are extracted at candidate spatio-temporal keypoints (STKP) and a HOPC quality factor is defined to rank the STKPs. Only high quality STKPs are retained. All points within the adaptive spatio-temporal support volume of each STKP are aligned along the eigenvectors of the spatial support around STKP. Thus the support volume is aligned with a local object centered coordinate basis and extracting HOPC, or any other feature, at the STKP will be view invariant. See Section 4.2 for details. Since humans may perform the same action at different speeds, to achieve speed invariance, we propose automatic temporal scale selection by minimizing the eigenratios over a varying temporal window size. The main contributions of this paper include:

– A HOPC descriptor for 3D pointclouds.

- A spatio-temporal key-point (STKP) detector and a view invariant descriptor.
- A technique for speed normalization of actions.

Moreover, we introduce a new 3D action dataset which has scale variations of subjects and viewpoint variations. It contains thirty actions which is larger number than any existing 3D action dataset. This dataset will be made public. Experimental comparison on four datasets, including three benchmark ones [13, 19, 32], with eight state-of-the-art methods [4, 10, 19, 21, 31, 32, 35, 36] shows the efficacy of our algorithms. Data and code of our technique are available [1].

## 2 Related Work

Based on the input data, human action recognition methods can be divided into three categories including RGB based, skeleton-based and depth based methods. In RGB videos, in order to recognize actions across viewpoint changes, mostly view independent representations are proposed such as view invariant spatio-temporal features [2, 20, 22, 23, 25, 33]. Some methods infer the 3D scene structure and use geometric transformations to achieve view invariance [5, 9, 15, 26, 39]. Another approach is to find a view independent latent space [7, 8, 12, 14] in which features extracted from the actions captured at different view points are directly comparable. Our proposed approach also falls in this category. However, our approach is only for 3D pointclouds captured by depth sensors. To the best of our knowledge, we are the first to propose cross-view action recognition using 3D pointclouds. We propose to normalize the spatio-temporal support volume of each candidate keypoint in the 3D pointcloud such that the feature extracted from the normalized support volume becomes view independent.

In skeleton based methods, 3D joint positions are used for action recognition. Multi-camera motion capture (MoCap) systems [3] have been used for human action recognition, but such special equipment is marker-based and expensive. Moreover, due to the different quality of the motion data, action recognition methods designed for MoCap are not suitable for 3D pointcloud sequences which is the focus of this paper [32].

On the other hand, some methods [36, 31, 37] use the human joint positions extracted by the OpenNI tracking framework (OpenNI) [24] as interest points. For example, Yang and Tian [37] proposed pairwise 3D joint position differences in each frame and temporal differences across frames to represent an action. Since 3D joints cannot capture all the discriminative information, the action recognition accuracy is compromised. Wang et al. [32] extended the previous approach by computing the histogram of occupancy pattern of a fixed region around each joint in a frame. In the temporal dimension, they used low frequency Fourier components as features and an SVM to find a discriminative set of joints. It is important to note that the estimated joint positions are not reliable and can fail when the human subject is not in an upright and frontal view position (e.g. lying on sofa) or when there is clutter around the subject.

Action recognition methods based on depth maps can be divided into holistic [19, 21, 13, 38, 29] and local approaches [32, 35, 11, 31]. Holistic methods use global features such as silhouettes and space-time volume information. For example, Li et al. [13] sampled boundary pixels from 2D silhouettes as a bag of features. Yang et al. [38] added temporal derivative of 2D projections to get Depth Motion Maps (DMM). Vieira et al.

[29] computed silhouettes in 3D by using the space-time occupancy patterns. Recently, Oreifej and Liu [19] extended histogram of oriented 3D normals [27] to 4D by adding time derivative. The gradient vector was normalized to unit magnitude and projected on a refined basis of 600-cell Polychrome to make histograms. The last component of normalized gradient vector was inverse of the gradient magnitude. As a result, information from very strong derivative locations, such as edges and silhouettes, may get suppressed [21]. The proposed HOPC descriptor is more informative than HON4D as it captures the spread of data in three principal directions. Thus, HOPC achieves more action recognition accuracy than exiting methods on three benchmark datasets.

Depth based local methods use local features where a set of interest points are extracted from the depth sequence and a feature descriptor is computed for each interest point. For example, Cheng et al. [4] used interest point detector proposed by Dollár et al. [11] and proposed a Comparative Coding Descriptor (CCD). Due to the presence of noise in depth sequences, simply extending color-based interest point detectors such as [6] and [11] may degrade the efficiency of these detectors [19].

Motion trajectory based action recognition methods[30, 34] are also not reliable in depth sequences [19]. Therefore, recent depth based action recognition methods resorted to alternative ways to extract more reliable interest points. Wang et al. [31] proposed Haar features to be extracted from each random subvolume. Xia and Aggarwal in [35] proposed a filtering method to extract spatio-temporal interest points. Their approach fails when the action execution speed is faster than the flip of the signal caused by the sensor noise. Both techniques are sensitive to viewpoint variations.

In contrast to previous interest point detection methods, the proposed STKP detector is robust to variations in action execution speed, sensor viewpoint and the spatial scale of the actor. Since the proposed HOPC descriptor is not strictly based on the depth derivatives, it is more robust to noise. Moreover, our methods do not require skeleton data which may be noisy or unavailable especially in the case of side views.

## 3 Histogram of Oriented Principal Component (HOPC)

Let $Q = \{Q_1, Q_2, \cdots, Q_t, \cdots, Q_{n_f}\}$ represent a sequence of 3D pointclouds captured by a 3D sensor, where $n_f$ denotes the number of frames (i.e. number of 3D pointclouds in the sequence) and $Q_t$ is the 3D pointcloud at time $t$. We make a spatio-temporal accumulated 3D pointcloud by merging the sequence of individual pointclouds in the time interval $[t-\tau, t+\tau]$. Consider a point $\mathbf{p} = (x_t \ y_t \ z_t)^\top, 1 \leq t \leq n_f$ in $Q_t$. We define the spatio-temporal support of $\mathbf{p}$, $\Omega(\mathbf{p})$, as the 3D points which are in a sphere of radius $r$ centered at $\mathbf{p}$ (Fig. 2). We propose a point descriptor based on the eigenvalue decomposition of the scatter matrix $C$ of the points $\mathbf{q} \in \Omega(\mathbf{p})$:

$$C = \frac{1}{n_p} \sum_{\mathbf{q} \in \Omega(\mathbf{p})} (\mathbf{q} - \mu)(\mathbf{q} - \mu)^\top, \text{where } \mu = \frac{1}{n_p} \sum_{\mathbf{q} \in \Omega(\mathbf{p})} \mathbf{q}, \tag{1}$$

and $n_p = |\Omega(\mathbf{p})|$ denotes the number of points in the spatio-temporal support of $\mathbf{p}$. Performing PCA on the scatter matrix $C$ gives us $CV = EV$, where $E$ is a diagonal matrix of the eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3$, and $V$ contains three orthogonal eigenvectors

$[\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3]$ arranged in the order of decreasing magnitude of their associated eigenvalues. We propose a new descriptor, the Histogram of Oriented Principal Components (HOPC), by projecting each eigenvector onto $m$ directions obtained from a *regular m-sided polyhedron*. We use $m = 20$ to make a *regular icosahedron* which is composed of 20 *regular pentagonal* facets and each facet corresponds to a histogram bin. Let $U \in \mathbb{R}^{3\times m}$ be the matrix of the center positions $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_m$ of facets:

$$U = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_i, \cdots, \mathbf{u}_m] \tag{2}$$

For a *regular icosahedron* with center at the origin, these normalized vectors are

$$\left(\frac{\pm 1}{L_u}, \frac{\pm 1}{L_u}, \frac{\pm 1}{L_u}\right), \left(0, \frac{\pm\varphi^{-1}}{L_u}, \frac{\pm\varphi}{L_u}\right), \left(\frac{\pm\varphi^{-1}}{L_u}, \frac{\pm\varphi}{L_u}, 0\right), \left(\frac{\pm\varphi}{L_u}, 0, \frac{\pm\varphi^{-1}}{L_u}\right), \tag{3}$$

where $\varphi = (1+\sqrt{5})/2$ is the golden ratio, and $L_u = \sqrt{\varphi^2 + 1/\varphi^2}$ is the length of vector $\mathbf{u}_i, 1 \le i \le m$. The eigenvectors are basically directions of maximum variance of the points in 3D space. Thus, they have a $180^\circ$ ambiguity. To overcome this problem, we consider the distribution of vector directions and their magnitudes within the support volume of $\mathbf{p}$. We determine the sign of each eigenvector $\mathbf{v}_j$ from the sign of the inner products of $\mathbf{v}_j$ and all vectors within the support of $\mathbf{p}$:

$$\mathbf{v}_j = \mathbf{v}_j . sign\left(\sum_{\mathbf{q}\in\Omega(\mathbf{p})} sign(\mathbf{o}^\top \mathbf{v}_j)(\mathbf{o}^\top \mathbf{v}_j)^2\right) \tag{4}$$

where $\mathbf{o} = \mathbf{q} - \mathbf{p}$ and the $sign$ function returns the sign of an input number. Note that the squared projection ensures the suppression of small projections, which could be due to noise. If the signs of eigenvectors $\mathbf{v}_1, \mathbf{v}_2$, and $\mathbf{v}_3$ disagree i.e. $\mathbf{v}_1 \times \mathbf{v}_2 \neq \mathbf{v}_3$, we switch the sign of the eigenvector whose $|\sum_{w=1}^{n_p} sign(\mathbf{o}_w^\top \mathbf{v}_j)(\mathbf{o}_w \mathbf{v}_j)^2|$ value is the smallest. We then project each eigenvector $\mathbf{v}_j$ onto $U$ to give us:

$$\mathbf{b}_j = U^\top \mathbf{v}_j \in \mathbb{R}^m, \textbf{ for } 1 \le j \le 3. \tag{5}$$

In case $\mathbf{v}_j$ is perfectly aligned with $\mathbf{u}_i \in U$, it should vote into only $i^{\text{th}}$ bin. However, all $\mathbf{u}_i$'s are not orthogonal, therefore $\mathbf{b}_j$ will have non-zero projection in other bins as well. To overcome this effect, we quantize the projection of $\mathbf{b}_j$. For this purpose, a threshold value $\psi$ is computed by projecting any two *neighbouring* vectors $\mathbf{u}_k$ and $\mathbf{u}_l$,

$$\psi = {\mathbf{u}_k}^\top \mathbf{u}_l = \frac{\varphi + \varphi^{-1}}{{L_u}^2}, \quad \mathbf{u}_k, \mathbf{u}_l \in U. \tag{6}$$

Note that for any $\mathbf{u}_k \in U$, we can find a $\mathbf{u}_l \in U$ such that $\psi = (\varphi + \varphi^{-1})/{L_u}^2$. The quantized vector is given by

$$\hat{\mathbf{b}}_j(z) = \begin{cases} 0 & \text{if } \mathbf{b}_j(z) \le \psi \\ \mathbf{b}_j(z) - \psi & \text{otherwise,} \end{cases}$$

where $1 \le z \le m$. We define $\mathbf{h}_j$ to be $\hat{\mathbf{b}}_j$ scaled by the corresponding eigenvalue $\lambda_j$,

$$\mathbf{h}_j = \frac{\lambda_j \cdot \hat{\mathbf{b}}_j}{||\hat{\mathbf{b}}_j||_2} \in \mathbb{R}^m, \textbf{ for } 1 \le j \le 3. \tag{7}$$

We concatenate the histograms of oriented principal components of all three eigenvectors in decreasing order of their eigenvalues to form a descriptor of point $\mathbf{p}$:

$$\mathbf{h_p} = [\mathbf{h}_1^\top \ \mathbf{h}_2^\top \ \mathbf{h}_3^\top]^\top \in \mathbb{R}^{3m}. \tag{8}$$

The spatio-temporal HOPC descriptor at point $\mathbf{p}$ encodes information from both shape and motion in the support volume around it. Since the smallest principal component of the local surface is in fact the total least squares estimate of the surface normal [18], our descriptor, which inherently encodes the surface normal, is more robust to noise than gradient-based surface normal used in [27, 19]. Using this descriptor, we propose two different action recognition algorithms in the following section.

## 4 Action Recognition

We propose a holistic and a local approach for human action recognition. Our holistic method is suitable for actions under occlusions, more inter-class similarities of local motions, and where the subjects do not change their spatial locations. On the other hand, our local method is more suitable for cross-view action recognition and in cases where the subjects change their spatial locations.

### 4.1 Action Recognition with Holistic HOPC

A sequence of 3D pointclouds is divided into $\gamma = n_x \times n_y \times n_t$ spatio-temporal cells along $X$, $Y$, and $T$ dimensions. We use $c_s$, where $s = 1 \cdots \gamma$, to denote the $s^{\text{th}}$ cell. The spatio-temporal HOPC descriptor $\mathbf{h_p}$ in (8) is computed for each point $\mathbf{p}$ within the sequence. The cell descriptor $\mathbf{h}_{c_s}$ is computed by accumulating $\mathbf{h}_{c_s} = \sum_{p \in c_s} \mathbf{h_p}$ and then normalizing $\mathbf{h}_{c_s} \leftarrow \mathbf{h}_{c_s} / ||\mathbf{h}_{c_s}||_2$. The final descriptor $\mathbf{h}_v$ for the given sequence is a concatenation of $\mathbf{h}_{c_s}$ obtained from all the cells:$\mathbf{h}_v = [\mathbf{h}_{c_1}^\top \ \mathbf{h}_{c_2}^\top \ ... \ \mathbf{h}_{c_s}^\top \ ... \ \mathbf{h}_{c_\gamma}^\top]^\top$. We use $\mathbf{h}_v$ as the holistic HOPC descriptor and use SVM for classification.

**Computing a Discriminative Cell Descriptor:** The HOPC descriptor is highly correlated to the order of eigenvalues of the spatio-temporal support volume around $\mathbf{p}$. Therefore, for each point a pruning approach is introduced to eliminate the ambiguous eigenvectors of each point. For this purpose, we define two eigenratios:

$$\delta_{12} = \frac{\lambda_1}{\lambda_2}, \delta_{23} = \frac{\lambda_2}{\lambda_3}. \tag{9}$$

For 3D symmetrical surfaces, the values of $\delta_{12}$ or $\delta_{23}$ will be equal to 1. The principal components of symmetrical surfaces are ambiguous. To get a discriminative $\mathbf{h_p}$, the values of $\delta_{12}$ and $\delta_{23}$ must be greater than 1. However, to manage noise we choose a threshold value $\theta > 1 + \epsilon$, where $\epsilon$ is a margin and select only the discriminative eigenvectors as follows:

1. If $\delta_{12} > \theta$ and $\delta_{23} > \theta$: $\mathbf{h_p} = [\mathbf{h}_1^\top \ \mathbf{h}_2^\top \ \mathbf{h}_3^\top]^\top$.
2. If $\delta_{12} \leq \theta$ and $\delta_{23} > \theta$: $\mathbf{h_p} = [\mathbf{0}^\top \ \mathbf{0}^\top \ \mathbf{h}_3^\top]^\top$.
3. If $\delta_{12} > \theta$ and $\delta_{23} \leq \theta$: $\mathbf{h_p} = [\mathbf{h}_1^\top \ \mathbf{0}^\top \ \mathbf{0}^\top]^\top$.
4. If $\delta_{12} \leq \theta$ and $\delta_{23} \leq \theta$: In this case, we discard $\mathbf{p}$.

### 4.2 STKP: Spatio-Temporal Key-Point Detection

Consider a point $\mathbf{p} = (x_t\ y_t\ z_t)^\top$ within a sequence of 3D pointclouds. In addition to the spatio-temporal support volume around $\mathbf{p}$ defined in section 3, we further define a spatial only support volume around $\mathbf{p}$ as the 3D points of $Q_t$ that fall inside a sphere of radius $r$ centered at $\mathbf{p}$. Thus, we perform PCA on both the spatial and the spatio-temporal scatter matrices $C'$ and $C$.

Let $\lambda'_1 \geq \lambda'_2 \geq \lambda'_3$ and $\lambda_1 \geq \lambda_2 \geq \lambda_3$ represent the eigenvalues of the spatial $C'$ and spatio-temporal $C$ scatter matrix, respectively. We define the following ratios:

$$\delta'_{12} = \frac{\lambda'_1}{\lambda'_2},\ \delta'_{23} = \frac{\lambda'_2}{\lambda'_3},\ \delta_{12} = \frac{\lambda_1}{\lambda_2},\ \delta_{23} = \frac{\lambda_2}{\lambda_3}. \tag{10}$$

For a point to be identified as a potential keypoint, the condition $\{\delta_{12}, \delta_{23}, \delta'_{12}, \delta'_{23}\} > \theta$ must be satisfied. This process prunes ambiguous points and produces a subset of candidate keypoints. It reduces the computational burden of the subsequent steps. Let $\mathbf{h}'_\mathbf{p} \in \mathbb{R}^{3m}$ represent the spatial HOPC and $\mathbf{h}_\mathbf{p} \in \mathbb{R}^{3m}$ represent the spatio-temporal HOPC. A *quality* factor is computed at each candidate keypoint $\mathbf{p}$ as follows:

$$\eta_p = \frac{1}{2}\sum_{i=1}^{3m} \frac{(\mathbf{h}'_\mathbf{p}(i) - \mathbf{h}_\mathbf{p}(i))^2}{(\mathbf{h}'_\mathbf{p}(i) + \mathbf{h}_\mathbf{p}(i))}. \tag{11}$$

When $\mathbf{h}'_\mathbf{p} = \mathbf{h}_\mathbf{p}$, the *quality* factor has the minimum value of $\eta_p = 0$. It means that the candidate keypoint $\mathbf{p}$ has a stationary spatio-temporal support volume with no motion.

We define a locality as a sphere of radius $r'$ ( with $r' \ll r$) and a time interval $2\tau'+1$ (with $\tau' \leq \tau$). We sort the candidate STKPs according to their quality values and starting from the highest quality keypoint, all STKPs within its locality are removed. The same process is repeated on the remaining STKPs. Fig. 2 shows the steps of our STKP detection algorithm. Fig. 3-a shows the extracted STKPs from three different views for a sequence of 3D pointclouds corresponding to the *holding head* action.

### 4.3 View-Invariant Key-Point Descriptor

Let $\mathbf{p} = (x_t\ y_t\ z_t)^\top$ represent an STKP. All points within the spatio-temporal support volume of $\mathbf{p}$ i.e., $\Omega(\mathbf{p})$, are aligned along the eigenvectors of its spatial scatter matrix, $B = PV'$, where $P \in \mathbb{R}^{n_p \times 3}$ is a matrix of points within $\Omega(\mathbf{p})$ and $V' = [\mathbf{v}'_1\ \mathbf{v}'_2\ \mathbf{v}'_3]$ denotes the $3 \times 3$ matrix of eigenvectors of the spatial scatter matrix $C'$. Recall that the signs of these eigenvectors have a $180^\circ$ ambiguity. As mentioned earlier, we use the sign disambiguation method to overcome this problem. As a result, any feature (e.g. raw depth values or HOPC) extracted from the aligned spatio-temporal support volume around $\mathbf{p}$ will be view invariant.

In order to describe the points within the spatio-temporal support volume of keypoint $\mathbf{p}$, the spatio-temporal support of $\mathbf{p}$ is represented as a 3D hyper-surface in the 4D space $(X, Y, Z)$ and $T$. We fit a 3D hyper-surface to the aligned points within the spatio-temporal support volume of $\mathbf{p}$. A uniform $m_x \times m_y \times m_t$ grid is used to sample the hyper-surface and its raw values are used as the descriptor of keypoint $\mathbf{p}$.

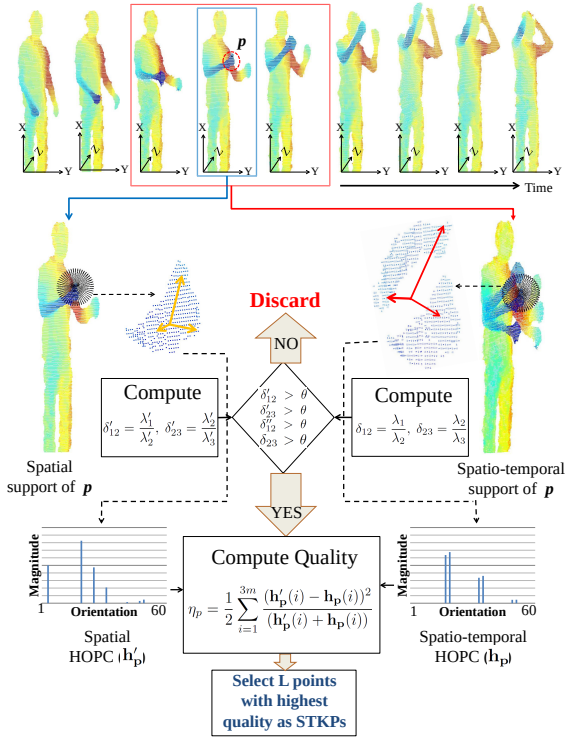



**Fig. 2.** STKP: Spatio-Temporal Key-Point detection algorithm

We use the bag-of-words approach to represent each 3D pointcloud sequence and build a codebook by clustering the keypoint descriptors using K-means. Codewords are defined by the cluster centers and descriptors are assigned to codewords using Euclidean distance. For classification, we use SVM with the histogram intersection kernel [16].

## 5 Adaptive Support Volume

So far we have used a fixed spatial ($r$) and temporal ($\tau$) support volume to detect and describe each keypoint $\mathbf{p}$. However, subjects can have different scales (in height and width) and perform actions with different speeds. Therefore, simply using a fixed spatial ($r$) and temporal ($\tau$) support volume is not optimal. Large values of $r$ and $\tau$ enable the proposed descriptors to encapsulate more information about shape and motion of a subject. However, this also increases sensitivity to occlusion and action speed.

A simple approach to finding the optimal spatial scale ($r$) for a STKP is based on the subject's height ($h_s$) e.g. $r = e \times h_s$, where $e$ is a constant that is empirically chosen to make a trade-off between descriptiveness and occlusion. This approach is unreliable and may fail when a subject touches the background or is not in an upright position. Several automatic spatial scale detection methods [28] have been proposed for 3D object recognition. In this paper, we use the automatic spatial scale detection method

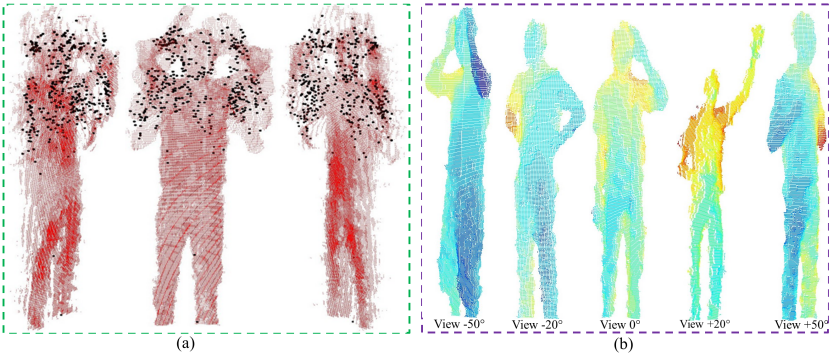



**Fig. 3.** (a)-STKPs projected onto $XY$ dimensions on top of all points within a sequence of 3D pointclouds corresponding to the *holding head* action (from three different views). Note that a large number of STKPs are detected only where movement is performed. (b)-Sample pointclouds at different views from the UWA3D Multiview Activity dataset

proposed by Mian et al. [17] to determine the optimal spatial scale for each keypoint. The optimal spatial scale ($r_b$) is selected as the one for which the ratio between the first two eigenvalues of the spatial support of a keypoint reaches a local maximum. Our results show that the automatic spatial scale selection [17] achieves the same accuracy as the fixed scale when the height ($h_s$) of each subject is available.

For temporal scale selection, most previous works [19, 35, 21, 29, 6] used a fixed number of frames. However, we propose automatic temporal scale selection to make our descriptor robust to action speed variations. Our proposed method follows the automatic spatial scale detection method by Mian et al. [17]. Let $Q = \{Q_1, Q_2, \cdots, Q_t, \cdots, Q_{n_f}\}$ represent a sequence of 3D pointclouds. For a point $\mathbf{p} = [x_t \;\; y_t \;\; z_t]^\top$, we start with points in $[Q_{t-\tau}, \cdots, Q_{t+\tau}]$ for $\tau = 1$ which are within its spatial scale $r$ (note that we assume $r$ as the optimal spatial scale for $\mathbf{p}$) and calculate the summation of ratio between the first two eigenvalues ($\lambda_2/\lambda_1$) and the last two eigenvalues ($\lambda_3/\lambda_2$) as:

$$A^{\tau}_{\mathbf{p}} = \frac{\lambda_2}{\lambda_1} + \frac{\lambda_3}{\lambda_2}, \tag{12}$$

where $\lambda_1 \geq \lambda_2 \geq \lambda_3$. This process continues for all $\tau = 1, \cdots, \Delta$ and the optimal temporal scale $\tau$ corresponding to the local minimum value of $A_{\mathbf{p}}$ found for point $\mathbf{p}$. A point which does not have a local minimum is not considered as a candidate keypoint.

## 6 Experiments

The proposed algorithms were evaluated on three benchmark datasets including MSRAction3D [13], MSRGesture3D [31], and ActionPairs3D [19]. We also developed a new "UWA3D Multiview Activity" dataset to evaluate the proposed cross-view action recognition algorithm. This dataset consists of 30 daily activities of ten subjects performed at different scales and viewpoints (Subsection 6.4). For our algorithms, we used $k =$

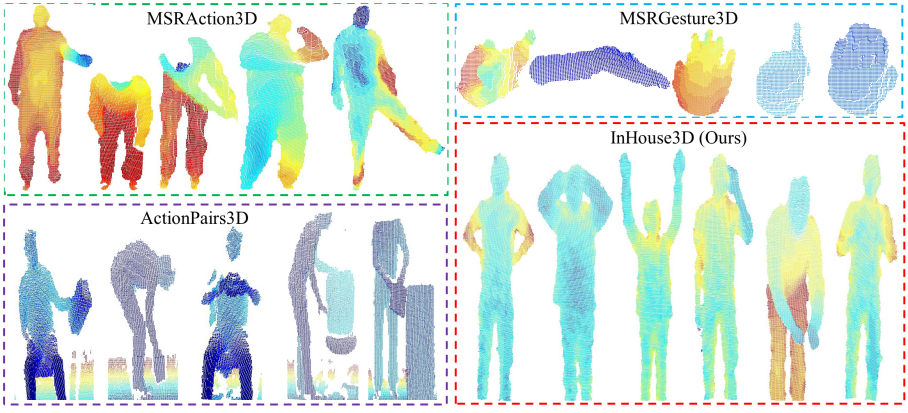



**Fig. 4.** Sample 3D pointclouds from the MSRAction3D, MSRGesture3D, ActionPairs3D, and UWA3D Multiview Activity datasets

$1000, \theta = 1.12$, $m_x = m_y = 20$ and $m_t = 3$ in all experiments. To test the performance of our holistic approach, each sequence of 3D pointclouds was divided into $6 \times 5 \times 3$ spatio-temporal cells along $X$, $Y$, and $T$ dimensions, respectively.

The performance of the proposed algorithms was compared with seven state-of-the-art methods including Histogram of Oriented Gradient (HOG3D) [10], Random Occupancy Pattern (ROP) [31], Histogram of 3D joints(HOJ3D) [36], Actionlet Ensemble [32], Histogram of 4D Oriented Normals (HON4D) [19], Depth Spatio-Temporal Interest Points (DSTIP) [35], and Histograms of Depth Gradient (HDG) [21]. The accuracy is reported from the original papers or from the authors' implementations of DSTIP [35], HDG [21], HOG3D [10], and HON4D [19]. The implementation of HOJ3D [36] is not available, therefore we used our own implementation.

### 6.1 MSRAction3D dataset

MSRAction3D dataset [13] consists of 20 actions each performed by 10 subjects 2-3 times (Fig. 4). The dataset is challenging due to high inter-action similarities. To test our holistic approach, we used five subjects for training and five for testing and repeated the experiments 252 folds exhaustively as proposed by [19]. To show the effectiveness of our automatic spatio-temporal scale selection, we used four different settings using fixed and varying values of $r$ and $\tau$. Table 1 compares our algorithms with existing state-of-the-art. Note that the proposed algorithm outperformed all techniques under all four settings. The maximum accuracy was achieved using constant $r$ and adaptive $\tau$. Adaptive $r$ did not improve results since there is little scale variation in this dataset. Note that HOJ3D [36], Moving Pose [40] and Actionlet [32] use skeleton data which is not always available.

We also evaluated our local method with automatic spatial and temporal scale selection and achieved $90.90\%$ accuracy (subjects $\{1,3,5,7,9\}$ used for training and the rest for testing). This is higher than $89.30\%$ of DSTIP [35] and $88.36\%$ of HON4D [19]. Note that DSTIP [35] only reported the accuracy of the best fold and used additional

**Table 1.** Accuracy comparison on MSRAction3D dataset. Mean $\pm$ STD is computed over 252 folds. Fold 5/5 means subjects $\{1,3,5,7,9\}$ used for training and the rest for testing. [a] Moving Pose [40] used different setting

| Method | Mean$\pm$STD | Max | Min | 5/5 |
|---|---|---|---|---|
| HOJ3D [36] | 63.55$\pm$5.23 | 75.91 | 44.05 | 75.80 |
| HOG3D [10] | 70.38$\pm$4.40 | 82.78 | 55.26 | 82.78 |
| ROP [31] | - | - | - | 86.50 |
| Moving Pose [40] | - | - | - | 91.70[a] |
| Actionlet [32] | - | - | - | 88.20 |
| HON4D [19] | 81.88$\pm$4.45 | 90.61 | 69.31 | 88.36 |
| DSTIP [35] | - | 89.30 | - | - |
| HDG [21] | 77.68$\pm$4.97 | 86.13 | 60.55 | 83.70 |
| **Holistic HOPC** | | | | |
| constant $r$, constant $\tau$ | 85.45$\pm$2.31 | 92.39 | 73.54 | 91.64 |
| adaptive $r$, constant $\tau$ | 84.78$\pm$2.89 | 91.64 | 72.41 | 90.90 |
| constant $r$, adaptive $\tau$ | 86.49$\pm$2.28 | 92.39 | 74.36 | 91.64 |
| adaptive $r$, adaptive $\tau$ | 85.01$\pm$2.44 | 92.39 | 72.94 | 91.27 |

steps such as mining discriminative features which can be applied to improve the accuracy of any descriptor. We did not include such steps in our method.

### 6.2 MSRGesture3D dataset

The MSRGesture3D dataset [31] contains 12 American sign language gestures performed 2-3 times by 10 subjects. For comparison with previous techniques, we use the leave-one-subject-out cross validation scheme proposed by [31]. Because of the absence of full body subjects (only hands are visible), we evaluate our methods in two settings only. Table 2 compares our method to existing state-of-the-art methods excluding HOJ3D [36] and Actionlet [32] since they require 3D joint positions which are not present in this dataset. Note that both variants of our method outperform all techniques by a significant margin achieving an average accuracy of $96.23\%$ which is $3.5\%$ higher than the nearest competitor HDG [21]. We also tested our local method with automatic spatial and temporal scale selection and obtained an accuracy of $93.61\%$.

### 6.3 ActionPairs3D dataset

The ActionPairs3D dataset [19] consists of depth sequences of six pairs of actions (Fig. 4) performed by 10 subjects. This dataset is challenging as each action pair has similar motion and shape. We used half of the subjects for training and the rest for testing as recommended by [19] and repeated the experiments 252 folds. Table 3 compares the proposed holistic HOPC descriptor in two settings with existing state-of-the-art methods. Our algorithms outperformed all techniques with $2.23\%$ improvement over the nearest competitor. Adaptive $\tau$ provides better improvement on this dataset compared to the previous two. We also evaluated our local method with automatic spatial

**Table 2.** Comparison with state-of-the-art methods on MSRGesture3D dataset

| Method | Mean±STD | Max | Min |
|---|---|---|---|
| HOG3D [10] | 85.23±12.12 | 100 | 50.00 |
| ROP [31] | 88.50 | - | - |
| HON4D [19] | 92.45±8.00 | 100 | 75 |
| HDG [21] | 92.76±8.80 | 100 | 77.78 |
| **Holistic HOPC** | | | |
| adaptive $r$, constant $\tau$ | 95.29±6.24 | 100 | 83.67 |
| adaptive $r$, adaptive $\tau$ | 96.23±5.29 | 100 | 88.33 |

**Table 3.** Accuracy comparisons on the ActionPairs3D dataset. Mean±STD are computed over 252 folds. 5/5 means subjects $\{6,7,8,9,10\}$ used for training and the rest for testing

| Method | Mean±STD | Max | Min | 5/5 |
|---|---|---|---|---|
| HOJ3D [36] | 63.81±5.94 | 67.22 | 50.56 | 66.67 |
| HOG3D [10] | 85.76±4.66 | 85.56 | 65.00 | 82.78 |
| Actionlet [32] | - | - | - | 82.22 |
| HON4D [19] | 96.00±1.74 | 100 | 91.11 | 96.67 |
| **Holistic HOPC** | | | | |
| constant $r$, constant $\tau$ | 97.15±2.21 | 100 | 88.89 | 97.22 |
| constant $r$, adaptive $\tau$ | 98.23±2.19 | 100 | 88.89 | 98.33 |

and temporal scale selection and obtained $98.89\%$ accuracy using subjects $\{6.7.8.9.10\}$ for training and the rest for testing.

### 6.4 UWA3D Multiview Activity dataset

We collected a new dataset using the Kinect to emphasize three factors: (1) Scale variations between subjects. (2) View-point variations. (3) All actions were performed in a continuous manner with no breaks or pauses. Thus, the start and end positions of body for the same actions are different. Our dataset consists of 30 activities performed by 10 human subjects of varying scales: *one hand waving, one hand Punching, sitting down, standing up, holding chest, holding head, holding back, walking, turning around, drinking, bending, running, kicking, jumping, moping floor, sneezing, sitting down(chair), squatting, two hand waving, two hand punching, vibrating, falling down, irregular walking, lying down, phone answering, jumping jack, picking up, putting down, dancing,* and *coughing* (Fig. 4). To capture depth videos from front view, each subject performed two or three random permutations of the 30 activities in a continuous manner. For cross-view action recognition, 5 subjects performed 15 activities from 4 different side views (see Fig. 3-b). We organized our dataset by segmenting the continuous sequences. The dataset is challenging due to self-occlusions and high similarity. For example, *drinking* and *phone answering* actions have very similar motion and only the hand location in these actions is slightly different. As another example, *lying down*

**Table 4.** Accuracy comparison on the UWA3D Activity dataset for same-view action recognition

| Method | Mean±STD | Max | Min |
|---|---|---|---|
| HOJ3D [36] | 48.59±5.77 | 58.70 | 28.93 |
| HOG3D [10] | 70.09±4.40 | 82.78 | 51.60 |
| HON4D [19] | 79.28±2.68 | 88.89 | 70.14 |
| HDG [21] | 75.54±3.64 | 85.07 | 61.90 |
| **Holistic HOPC** | | | |
| constant $r$, constant $\tau$ | 83.77±3.09 | 92.18 | 74.67 |
| constant $r$, adaptive $\tau$ | 84.93±2.75 | 93.11 | 74.67 |

and *falling down* actions have very similar motion, but the speed of action execution is different. Moreover, some actions such as: *holding back, holding head,* and *answering phone* have self-occlusions. The videos were captured at 30 frames per second at a spatial resolution of $640 \times 480$.

We evaluate our proposed methods in the same-view, and cross-view action recognition settings. The holistic approach is used to classify actions captured from the same view and the local approach is used for cross-view action recognition where the training videos are captured from front view and the test videos from side views.

**Same-view Action Recognition** We selected half of the subjects as training and the rest as testing and evaluated our holistic method in two settings: (1) constant $r$, constant $\tau$, (2) constant $r$, adaptive $\tau$. Table 4 compares our methods with existing state of the art. Both variants of our algorithm outperform all methods achieving a maximum of $84.93\%$ accuracy. The adaptive $\tau$ provides minor improvement because there is no explicit action speed variation in the dataset. To further test the robustness of our temporal scale selection (adaptive $\tau$) to action speed variations we use depth videos of actions performed by half of the subjects captured at 30 frames per second as training data and depth videos of actions performed by the remaining subjects captured at 15 frames per second as test data. The average accuracy of our method using automatic temporal scale selection was $84.64\%$ which is higher than $81.92\%$ accuracy achieved by our method using constant temporal scale and the $76.43\%$ accuracy achieved by HON4D. Next, we swap the frame rates of the test and training data. The average accuracy of our method using automatic temporal scale selection was $84.70\%$ which is higher than $81.01\%$ accuracy achieved by our method using constant temporal scale. The accuracy of HON4D was $75.81\%$ in this case.

**Cross-view Action Recognition** In order to evaluate the STKP detector and HOPC descriptor for cross-view action recognition, we used front views of five subjects as training and side views of the remaining five subjects as test. Table 5 compares our method with existing state-of-the- art holistic and local methods for cross-view action recognition. Note that the performance of all other methods degrades when the subjects perform actions at different viewing angles. This is not surprising as existing methods assume that actions are observed from the same viewpoint i.e. frontal. For example,

**Table 5.** Cross-view action recognition on the UWA3D Multiview Activity dataset. Depth sequences of five subjects at $0^o$ are used for training and the remaining subjects at $0^o$ and 4 different side-views are used for testing. Average accuracy is computed only for the cross-view scenario

| Method | View angle $0^\circ$ | $-25^\circ$ | $+25^\circ$ | $-50^\circ$ | $+50^\circ$ | Average |
|---|---|---|---|---|---|---|
| **Holistic Methods** | | | | | | |
| HON4D [19] | 86.55 | 62.22 | 60.00 | 35.56 | 37.78 | 48.89 |
| HDG [21] | 79.13 | 60.00 | 64.44 | 33.33 | 35.56 | 48.33 |
| **Local Methods** | | | | | | |
| HOJ3D [36] | 63.34 | 60.00 | 62.22 | 37.78 | 40.00 | 50.00 |
| DSTIP+DCSF [35] | 80.80 | 66.67 | 71.11 | 35.56 | 40.00 | 53.33 |
| STKP+hyper-surface fitting | 87.39 | 81.33 | 82.67 | 71.11 | 71.11 | 76.56 |
| STKP+HOPC | **91.79** | **86.67** | **88.89** | **75.56** | **77.78** | **82.23** |

HON4D achieved $86.55\%$ accuracy when the training and test samples were in the same view (frontal). The average accuracy of HON4D dropped to $48.89\%$ when the training samples were captured from front view and the test samples were captured from four different side views. We also observed that the performance of existing methods did not degrade only for actions like *standing up, sitting down*, and *turning around*. This is due to the distinctness of these actions regardless of the viewpoint.

We test two variants of our method. First, we apply our STKP detector on 3D point-cloud sequences and use the raw values of fitted hyper-surface as features. The average accuracy obtained over the four different side views ($\pm 25^\circ$ and $\pm 50^\circ$) was $76.56\%$ in this case. Next, we use the STKP detector combined with the proposed HOPC descriptor. This combination achieved the best average accuracy i.e. $82.23\%$. Comparison with other methods and the accuracy of each method on different side views are shown in Table 5. These experiments demonstrate that our STKP detector in conjunction with HOPC descriptor significantly outperforms state-of-the-art methods for cross-view as well as same-view action recognition.

## 7 Conclusion

Performance of current 3D action recognition techniques degrades in the presence of viewpoint variations across the test and the training data. We proposed a novel technique for action recognition which is more robust to action speed and viewpoint variations. A new descriptor, Histogram of Oriented Principal Components (HOPC), and a keypoint detector are presented. The proposed descriptor and detector were evaluated for activity recognition on three benchmark datasets. We also introduced a new multiview public dataset and showed the robustness of our proposed method to viewpoint variations.

## Acknowledgment

This research was supported by ARC Discovery Grant DP110102399.

## References


1. UWA3D Multiview Activity dataset and Histogram of Oriented Principal Components Matlab code. http://www.csse.uwa.edu.au/∼ajmal/code.html (2014)
2. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV (2005)
3. Campbell, L., Bobick, A.: Recognition of human body motion using phase space constraints. In: ICCV (1995)
4. Cheng, Z., Qin, L., Ye, Y., Huang, Q., Tian, Q.: Human daily action analysis with multi-view and color-depth data. In: ECCVW (2012)
5. Darrell, T., Essa, I., Pentland, A.: Task-specific gesture analysis in real-time using interpolated views. In: PAMI (1996)
6. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: ICCV (2005)
7. Farhadi, A., Tabrizi, M.K.: Learning to recognize activities from the wrong view point. In: ECCV (2008)
8. Farhadi, A., Tabrizi, M.K., Endres, I., Forsyth, D.A.: A latent model of discriminative aspect. In: ICCV (2009)
9. Gavrila, D., Davis, L.: 3D model-based tracking of humans in action: a multi-view approach. In: CVPR (1996)
10. Klaeser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: BMVC (2008)
11. Laptev, I.: On space-time interest point. In: IJCV (2005)
12. Li, R.: Discriminative virtual views for cross-view action recognition. In: CVPR (2012)
13. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3D points. In: CVPRW (2010)
14. Liu, J., Shah, M., Kuipersy, B., Savarese, S.: Cross-view action recognition via view knowledge transfer. In: CVPR (2011)
15. Lv, F., Nevatia, R.: Single view human action recognition using key pose matching and viterbi path searching. In: CVPR (2007)
16. Maji, S., Berg, A.C., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: CVPR (2008)
17. Mian, A., Bennamoun, M., Owens, R.: On the repeatability and quality of keypoints for local feature-based 3D object retrieval from cluttered scenes. In: IJCV (2010)
18. Mitra, N.J., Nguyen, A.: Estimating surface normals in noisy point clouds data. In: SCG (2003)
19. Oreifej, O., Liu, Z.: HON4D: histogram of oriented 4D normals for activity recognition from depth sequences. In: CVPR (2013)
20. Parameswaran, V., Chellappa, R.: View invariance for human action recognition. In: IJCV (2006)
21. Rahmani, H., Mahmood, A., Huynh, D.Q., Mian, A.: Real time human ation recognition using histograms of depth gradients and random decision forests. In: WACV (2014)
22. Rao, C., Yilmaz, A., Shah, M.: View-invariant representation and recognition of actions . In: IJCV (2002)
23. Seitz, S., Dyer, C.: View-invariant analysis of cyclic motion. In: IJCV (1997)
24. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR (2011)
25. Syeda-Mahmood, T., Vasilescu, A., Sethi, S.: Recognizing action events from multiple viewpoints. In: IEEE Workshop on Detection and Recognition of Events in Video (2001)

26. Syeda-Mahmood, T., Vasilescu, A., Sethi, S.: Action recognition from arbitrary views using 3D exemplars. In: ICCV (2007)
27. Tang, S., Wang, X., Lv, X., Han, T., Keller, J., He, Z., Skubic, M., Lao, S.: Histogram of oriented normal vectors for object recognition with a depth sensor. In: ACCV (2012)
28. Timbari, F., Stefano, L.D.: Performance evaluation of 3D keypoint detectors. In: IJCV (2013)
29. Vieira, A.W., Nascimento, E., Oliveira, G., Liu, Z., Campos, M.: STOP: space-time occupancy patterns for 3D action recognition from depth map sequences. In: CIARP (2012)
30. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: CVPR (2011)
31. Wang, J., Liu, Z., Chorowski, J., Chen, Z., Wu, Y.: Robust 3D action recognition with random occupancy patterns. In: ECCV (2012)
32. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: CVPR (2012)
33. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. In: CVIU (2006)
34. Wu, S., Oreifej, O., Shah, M.: Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In: ICCV (2011)
35. Xia, L., Aggarwal, J.: Spatio-temporal depth cuboid similarity feature for activity recongition using depth camera. In: CVPR (2013)
36. Xia, L., Chen, C.C., Aggarwal, J.K.: View invariant human action recognition using histograms of 3D joints. In: CVPRW (2012)
37. Yang, X., Tian, Y.: EigenJoints-based action recognition using naive bayes nearest neighbor. In: CVPRW (2012)
38. Yang, X., Zhang, C., Tian, Y.: Recognizing actions using depth motion maps-based histograms of oriented gradients. In: ACM ICM (2012)
39. Yilmaz, A., Shah, M.: Action sketch: a novel action representation. In: CVPR (2005)
40. Zanfir, M., Leordeanu, M., Sminchisescu, C.: The moving pose: an efficient 3D kinematics descriptor for low-latency action recognition and detection. In: ICCV (2013)