Evaluating spatiotemporal interest point features for depth-based action recognition

Yu Zhu 1, Wenbin Chen 1, Guodong Guo *

Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506, United States

A R T I C L E   I N F O

Article history:
Received 11 August 2013
Received in revised form 6 January 2014
Accepted 2 April 2014
Available online 16 April 2014

Keywords:
Action recognition
Spatiotemporal interest point (STIP)
Detectors
Descriptors
STIP features
RGB-D sensor
Evaluation
STIP feature refinement
Feature fusion

A B S T R A C T

Human action recognition has lots of real-world applications, such as natural user interface, virtual reality, intelligent surveillance, and gaming. However, it is still a very challenging problem. In action recognition using the visible light videos, the spatiotemporal interest point (STIP) based features are widely used with good performance. Recently, with the advance of depth imaging technology, a new modality has appeared for human action recognition. It is important to assess the performance and usefulness of the STIP features for action analysis on the new modality of 3D depth map. In this paper, we evaluate the spatiotemporal interest point (STIP) based features for depth-based action recognition. Different interest point detectors and descriptors are combined to form various STIP features. The bag-of-words representation and the SVM classifiers are used for action learning. Our comprehensive evaluation is conducted on four challenging 3D depth databases. Further, we use two schemes to refine the STIP features, one is to detect the interest points in RGB videos and apply to the aligned depth sequences, and the other is to use the human skeleton to remove irrelevant interest points. These refinements can help us have a deeper understanding of the STIP features on 3D depth data. Finally, we investigate a fusion of the best STIP features with the prevalent skeleton features, to present a complementary use of the STIP features for action recognition on 3D data. The fusion approach gives significantly higher accuracies than many state-of-the-art results.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Human actions convey a significant amount of information for human interaction with the environment, human-to-human communication and human-to-machine interaction. Human action recognition is a very active research topic in computer vision, aiming to automatically recognize and interpret ongoing human actions. The ability to recognize complex human actions from videos enables the construction of several important applications such as natural user interfaces, virtual reality, intelligent surveillance and gaming [1,2].

Although human action recognition is very important for many real-world applications, it is still a challenging problem. A number of methods have been proposed to solve the action recognition problem [2]. Among various methods, the spatiotemporal interest point (STIP) based features have shown good performance for action recognition in RGB videos [3].

Very recently, depth imaging technology has made a significant progress, which brings a broader scope for human action recognition. Using a consumer depth sensor, e.g., the Kinect [4], depth information can be captured simultaneously with the RGB videos. Moreover, from the depth maps the geometric positions of skeleton points can also be detected effectively [4]. As a result, the depth data provides a promising modality for action recognition.

In traditional RGB video-based action recognition, several spatiotemporal features have been proposed to characterize human actions using local motions in a space-time volume. Local features possess many advantages, e.g., it can avoid possible problems caused by inaccurate segmentation or partial occlusions. In the literature, many spatiotemporal feature detectors [5–8] and descriptors [9–12] have been proposed and shown promising performance for action recognition in RGB videos. However, it has not been well studied yet on whether these spatiotemporal interest point (STIP) features can be useful or not for depth-based action recognition.

In this paper, we perform a comprehensive evaluation of different spatiotemporal interest point features for depth-based human action recognition. In particular, three interest point detectors and six local descriptors are adopted, in total there are 14 different detector/descriptor combinations adopted for the evaluation. Experiments are conducted on four challenging depth action databases with the same experimental setup for each feature. Besides, we also extend the capability of using spatiotemporal features by utilizing the corresponding RGB videos, and the skeleton joint positions, in order to have a deep understanding of the STIP features on depth data. Two different interest point refinement approaches are examined. Moreover, a feature-level fusion method is presented to combine the best spatiotemporal features on each database with the skeleton joint features. From the experimental
results and comparisons with the state-of-the-art approaches for depth-based action recognition, we show the usefulness of spatiotemporal features for action recognition in depth videos.

The rest of the paper is organized as follows: the related work on depth-based action recognition is reviewed in Section 2. Different spatiotemporal interest point features are introduced in Section 3. Four different depth action/activity databases are presented in Section 4. Experiments are conducted and presented in Section 5. Two STIP refinement approaches are introduced and evaluated experimentally. A fusion of the best STIP features with skeleton features is shown in Section 6. Finally, we draw conclusions.

2. Related work on depth-based action recognition

The depth sensors offer several advantages over traditional video cameras, e.g., working in low light conditions, giving a real 3D measure invariant to surface color and texture, and resolving silhouette ambiguities in pose [4]. Depth sensors can significantly simplify the task of background subtraction and human detection. Because of the advantages, the depth sensors, e.g., the Kinect, have attracted researchers’ attentions from many areas including 3D modeling, object recognition, gesture analysis, etc. Recently, action analysis and recognition in depth videos have become a very active topic. In this quite novel area, different approaches have been proposed. Here we give a brief overview of the methods for depth-based action recognition.

Li et al. [13] proposed a sampling of 80 representative 3D points to describe a salient posture. In order to select the representative points, each depth map was projected onto three orthogonal Cartesian planes: xy, xz and yz, and then a specified number of 2D points were sampled at equal distance along the contours of the projected depth data. An action graph was used to model the dynamics of actions. Their method has smaller error rates than using 2D silhouettes.

Xia et al. [14] proposed to use histograms of 3D joint locations (HZ3D) for action recognition. In order to be view invariant, they aligned the spherical coordinates with the person’s specific direction. The hip center joint served as the center of the coordinate system. By projecting the vector from left–hip center to the right–hip center, the horizontal reference vector was obtained. The zenith reference vector passes through the coordinate center and is perpendicular to the ground plane. According to different joint’s contribution to the body motion, they chose 9 joints to compute the 3D spatial histogram by partitioning the 3D space into 84 bins. After that, the LDA was performed to extract the dominant features, so that each frame will have a n − 1 dimensional feature vector, where n is the number of classes. The K-means clustering was performed to represent each posture as a visual word. A discrete SVM was trained for action recognition.

Vieira et al. [15] proposed the Space–Time Occupancy Patterns (STOP) to represent sequences of depth maps. In their representation, the space and time axes were divided into multiple segments so that each depth map sequence was embedded in multiple 4D grids. They computed occupancy feature in each cell. After that, they employed a Nearest Neighbor classifier based on the cosine distance for action recognition.

Yang and Tian [16] combined static posture, motion property, and overall dynamics to form an action feature descriptor called Eigenjoints. In order to remove noisy frames and reduce computational cost, they performed informative frame selection based on Accumulated Motion Energy (AME). A non-parametric Naïve–Bayes–Nearest–Neighbor (NBNN) classifier was used for action classification.

In order to make skeleton representation invariant to sensor orientation and global translation of the body, Miranda et al. [17] proposed a pose descriptor vector in a torso-based coordinate system. A predefined key pose set was used to build SVM classifiers. Because each gesture can be viewed as a sequence of key poses, a decision forest was used to search for key pose sequences. In recognition stage, the key pose classifiers can recognize key poses performed by the user and then determine the corresponding gesture class.

Yang et al. [18] proposed to generate three 2D Depth Motion Maps (DMM) from each 3D depth frame according to front, side, and top views. The HOG feature is computed from DMM to represent an action video. They used a linear SVM classifier to recognize actions.

In [19], Wang et al. extracted two features, pairwise relative positions and Local Occupancy Patterns at each joint. Each skeleton joint i has 3 coordinates Fi(t) = (xi(t), yi(t), zi(t)) at frame t, the pairwise relative position features are extracted for joint i as: pi = (pi[i] − j) = (pi − pj[l] ≠ j). In order to model the interaction between human subject and objects, they computed the LOP feature based on the 3D point cloud around a particular joint. After that, Fourier temporal pyramid was used to represent the temporal dynamics of the frame-level features. In order to deal with the errors of the skeleton tracking and better characterize the intra-class variations, they defined an actionlet as a conjunction structure on base features. One base feature is the Fourier pyramid feature of each joint. A data mining algorithm was used to find discriminative actionlets for action recognition.

In [20], Sung et al. used all three channels, i.e., RGB, depth and skeleton positions, for human activity recognition. They extracted hand position information, body pose features and motion from skeleton joints. For both RGB and depth images, they used the Histogram of Oriented Gradients (HOG) feature in two settings: One is to compute HOG in both the RGB and depth within the bounding box of the person. The other is to get the bounding boxes for the head, torso, left arm, and right arm, based on the skeleton locations, and compute the HOG in RGB and depth with each of the four bounding boxes. A two-layered maximum-entropy Markov model was trained to capture the hierarchies of human activities and transitions between sub-activities over time.

Wang et al. [21] proposed a semi-local feature called random occupancy patterns (ROP). A depth sequence is treated as a 4D volume. Given a subvolume, the ROP feature was computed as: oxyz = lq ∑ l=0 bwinlq, where lq = 1 if the point cloud has a point in the location q and lq = 0 otherwise. oxyz is a sigmoid normalization function: oxyz = 1 / 1 + e−xyz. Because the sizes of the 4D subvolume are extremely large and the features are highly redundant, a weighted sampling method was applied to reduce the complexity and obtain the discriminative features. They also utilized a sparse coding method to robustly encode those features. The SVM classifier was used for classification.

More recently, Oreifej et al. [22] represented the depth sequence using a histogram capturing the distribution of the surface normal orientation in 4D space of time, depth, and spatial coordinates (HON4D feature). A 600-cell polyhedral with 120 vertices was used to quantize the 4D space and represent possible directions of the 4D normals. The SVM classifier was used for action classification.

Koppula et al. [23] proposed to jointly model the human activities and object affordances as a Markov Random Field where the nodes represent objects and sub-activities, and the edges represent the relationships between object affordances, their relations with sub-activities, and their evolutions over time. In order to find atomic movements in an activity, they also performed temporal segmentation of the frames. They used a multi-class SVM classifier for action recognition.

Ni et al. [24] proposed the Depth-Layered Multi–Channel STIP (DLMC–STIP) and Three–Dimensional Motion History Images (3D–MHIs). For DLMC–STIP, after getting local feature descriptors in a video, they introduced a set of (M) depth layers L1 = [z1, z2, z3], L2 = [z4, z5, z6], ..., Lm = [zm, zm+1, zm+2], with lower and upper boundaries denoted as zm and zm+1 for the m-th depth layer, so a detected spatio-temporal interest point by Harris3D detector would be located in one specific layer. In this way they formed multi-channel histograms for feature description using the HOG/HOF descriptor. The 3D–MHIs are motion history images (MHIs), including both forward–DMHIs (fDMHIs) and backward–DMHIs (bDMHIs). The SVM classifiers were used for action recognition.
Zhao et al. [25] explored the combination of RGB channel and depth for action recognition. They extracted interest points from RGB videos. They proposed local depth pattern (LDP) to represent each local video volume at each interest point position extracted from visible light videos, and adopted to the corresponding depth videos. Given an interest point $p$, its local region is partitioned into $N_x \times N_y$ spatial cells. Each cell is of size $(S_x, S_y)$ pixels. For each cell, they computed an average depth value and then the difference of average depth values between every cell pair to form the LDP feature. For each interest point $p$, which can be detected by the Harris3D detector on either RGB video or depth sequence, the output feature vector can be denoted as $S_p = ((x, y, t), F)$, where $(x, y, t)$ denote the coordinates and time of interest point, and action feature $F$ could be obtained either by HOGHOF descriptor or LDP. They explored different combinations of RGB and depth map features and used the SVM classifiers.

Inspired by Dollar’s work on local features [6], Zhang et al. [26] developed a 4D local spatio-temporal feature which combines both intensity and depth information. They first applied separate filters along the 3D spatial dimensions and the temporal dimension to detect interest point. Then they computed and concatenated the intensity and depth gradients with a 4D hyper cuboid to obtain features for an action sequence. The Latent Dirichlet Allocation with Gibbs sampling was used as the classifier.

Also inspired by Dollar’s local interest point detector [6], Xia and Aggarwal [27] proposed a spatiotemporal interest point detector on depth map, which effectively eliminates the noise (‘value jumps’ and ‘holes’) appear on depth maps. They extended the cuboid detector [6] to the fourth dimension. A depth cuboid similarity descriptor is proposed to describe the local feature, based on the similarity between all pair of blocks in the 3D cuboid. Finally a feature selection process based on F-score is applied to generate the feature vector, and then used for action classification.

From the overview of related works on depth-based action recognition, we show that many of the approaches were motivated by the methods originally developed for RGB action recognition, e.g., motion history and the spatiotemporal interest point features. Although the STIP features prevail for analyzing color/intensity actions with good performance, only very limited types of STIP features were applied to depth-based action recognition. It has not been well studied yet on the performance of the typical STIP features on 3D depth actions. Thus it is important to evaluate the representative STIP features, so that a better understanding of the STIP features can be obtained for 3D depth-based action analysis. Our goal is to measure the usefulness of the STIP features for 3D action recognition, and build benchmark results of these features on several depth-based action databases.

In the following, we briefly introduce the STIP features that we used for the evaluation, and then present the databases and the evaluation results.

### 3.1. Interest points detectors

The Harris3D detector was proposed in [5]. It locates the spatiotemporal volumes with large variations along space and temporal directions in a video sequence. A spatiotemporal second-moment matrix is used to model a video sequence $f_{\vec{u}} = g(\cdot) \times \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t & L_x L_{xy} \\ L_x L_y & L_y^2 & L_y L_t & L_y L_{xy} \\ L_x L_t & L_y L_t & L_t^2 & L_t L_{xy} \\ L_x L_{xy} & L_y L_{xy} & L_t L_{xy} & L_{xy}^2 \end{pmatrix}$, where $g(\cdot)$ is a Gaussian function or weighting and $L$ is the convolution of $f$ with a spatiotemporal Gaussian derivative kernel. The interest point locations are determined by the local maxima of the response function $H = det(\mu) - k \cdot trace^{2}(\mu)$.

The cuboid [6] detector computes the interest point location by the local maxima of the response function $R$, which is defined as: $R = (I \ast g \ast h_{vol})^2 + (I \ast g \ast h_{vol})^2$, where $g$ is the 2D Gaussian smoothing kernel, $h_{vol}$ and $h_{vol}$ are a quadrature pair of 1D Gabor filter, which are computed by $h_{vol} = -cos(2\pi u\tau)e^{-\pi(\tau^2+\omega^2)}$ and $h_{vol} = -sin(2\pi u\tau)e^{-\pi(\tau^2+\omega^2)}$.

Willems et al. [9] proposed the Hessian detector, which measures the strength of each interest point using the Hessian matrix. The response function is defined as $S = |det(H)|$, where $H$ is the Hessian matrix.

### 3.2. Local feature descriptors

Given a set of interest point locations, various feature descriptors can be applied to characterize the local space–time content. Given the spatial scale $\sigma$ and temporal scale $\tau$ at each interest point location, a local volume is used to extract features.

Kläser et al. extended the histograms of oriented gradient (HOG) to HOG3D, which is the histogram of 3D gradient orientations. Integral volumes are computed for efficiency. HOG/HOF descriptor was proposed by Laptev et al. [30], using the combination of histogram of gradient (HoG) and histogram of optical flow (HoF) accumulated from the local volume.

The cuboid descriptor was proposed along with the cuboid detector in [6]. For each detected point $(x, y, t, \sigma, \tau)$, a feature descriptor is computed in a 3D patch centered at $(x, y, t)$. The gradient at each spatiotemporal location is computed within the cuboid and the histogram is computed as the feature vector. The PCA can be applied to reduce the dimensionality.

The extended SURF (ESURF) descriptor [9] was proposed with the Hessian detector, which is an extension of the SURF [32]. For each local volume, the feature vector is computed using the sum of uniformly sampled responses of Haar–wavelets along three directions.

We will evaluate the above three interest point detectors and six local descriptors for 3D action recognition. Although there exist some works using the STIP features for depth-based action recognition [25–27], only very limited types of STIP features were investigated. Through the evaluation of several representative STIP features on multiple depth databases, we will not only provide the benchmark results of STIP features on depth data, but also find the best, appropriate STIP features that may help to improve the accuracies significantly [33] for depth-based action recognition.
4. Databases

In order to perform a comprehensive evaluation, we conduct experiments on four different depth databases, which were captured under different scenarios and/or environments. The evaluation on these databases can provide a thorough test of various STIP features on depth data. Table 1 shows a brief description of the four depth-based action/activity databases. More details of these databases are given as follows.

4.1. MSR-Action3D dataset

MSR-Action3D dataset [13] was captured by a depth camera similar to the Kinect sensor. This dataset contains 20 actions, and each action was performed by 10 subjects three times. Two channels of data are provided: depth sequences at 15 frames per second (fps) with resolution of 640 × 480, and skeleton joint positions in each frame. The 20 actions are: high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, handclap, two hand wave, sideboxing, bend, forward kick, side kick, joggling, tennis swing, tennis serve, golf swing, and pick up & throw (see Fig. 1 for some example images).

4.2. MSRDailyActivity3D dataset

This dataset was collected for human daily activities by a Kinect device [19]. In total there are 16 activities in this dataset: drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lay down on sofa, walk, play guitar, stand up, and sit down. Each subject performed an activity twice, one “sitting on sofa” and the other “standing”. The total number of videos is 320. Three channels of data, i.e., RGB, depth and skeleton joint positions are provided in this dataset. See Fig. 2 for some examples of depth images in this dataset.

4.3. UTKinect-Action dataset

The action videos of the UTKinect-Action dataset [14], were collected by a single stationary Kinect with the distance ranges from 4 to 11 ft. There are totally 10 action classes performed by 10 subjects. Each subject performed each action twice. The RGB, depth and skeleton joint locations are synchronized and all three channels are provided. Some examples of depth images are shown in Fig. 3. The resolution of RGB images is 640 × 480, the depth image resolution is 320 × 240. The 10 action classes are: walk, sit down, stand up, pick up, carry, throw, push, pull, wave hands, and clap hands.

4.4. CAD-60 dataset

Cornell activity dataset-60 (CAD-60) [20], contains 60 RGB-D videos collected by a Kinect sensor with the distance ranges from 1.2 m to 3.5 m, the resolution of the depth sequences is 640 × 480, and captured at 15 fps. There are 4 different subjects and 12 different actions. The action videos were captured in five different locations, with 3 to 4 common activities performed at each location. The five locations are: office, kitchen, bedroom, bathroom and living room. Fig. 4 shows some example depth images from this dataset. All the RGB, depth and skeleton data are provided in this dataset.

5. Evaluations

We present the experimental settings in Section 5.1, the evaluation results for various combinations of detectors and descriptors in Section 5.2, and two STIP refinement approaches along with the corresponding results in Section 5.3.

5.1. Experimental settings

The bag-of-words representation is used for the spatiotemporal interest points. First, different STIP detectors are applied to the depth sequences. Given the detected locations, different local descriptors are used to characterize the space–time volume around each interest point. These local features are then quantized into visual words, so that a depth action sequence can be represented as a histogram of the visual words. In our evaluation, vocabularies are constructed using the K-means clustering technique. We empirically set the vocabulary size to be 200, 300, 850 and 1550, respectively, for the MSRDailyActivity3D, MSRAction3D, CAD-60 dataset and UTKinect-Action datasets, depending on the database size and empirical performance. After quantization, the histograms of visual words are used as the features for action classification. The multi-class support vector machines (SVMs) are used for action learning, with a linear kernel for the CAD-60 dataset and a $\chi^2$-kernel for the other three datasets, based on our empirical comparisons.
between different kernels. The $\chi^2$-kernel is defined by:

$$K(H_i, H_j) = \exp\left( -\frac{1}{2A} \sum_{n=1}^{V} \left( \frac{h_{in} - h_{jn}}{h_{in} + h_{jn}} \right)^2 \right),$$

where $H_i = \{h_{in}\}$ and $H_j = \{h_{jn}\}$ are the frequency histograms of the visual word occurrences, and $V$ is the vocabulary size. $A$ is the mean value of distances between all training samples.

For different feature representations, we utilize the implementations or source code provided by the authors, mostly with the default parameter settings, since some executable code cannot be modified. All the experiments were conducted on a 64-bit operating system DELL Optiplex 790 PC, with i7 3.4GHz CPU and 12G RAM.

Specifically, for the Harris3D detector, we used the original implementation with the default parameter settings: $k = 0.0005$, $\sigma^2 = [4, 8, 16, 32, 64, 128]$ and $\tau^2 = [2, 4]$. For the cuboid detector [6], we ran the authors’ implementation and the default scale values $\sigma = 2$, $\tau = 4$ were used in our evaluation. The UTKinect-Action dataset has typically shorter video clips, we used $\sigma = 2$, $\tau = 2$ for the cuboid detector. For the Hessian detector [9], the executable code was used with the default parameter setting.

For the HOG/HOF descriptor, we followed [30] and adopted the grid parameters $n_x = n_y = 3$, $n_t = 2$, $\sigma^2 = 4$ and $\tau^2 = 2$. For the HOG3D descriptor [29], we used the parameters $n_x = n_y = 5$, $n_t = 4$, $\sigma = 2$ and $\tau = 2$ for the UTKinect-Action dataset and $n_x = n_y = 2$, $n_t = 5$, $\sigma = 2$ and $\tau = 4$ for the other three datasets in our evaluation. For the cuboid descriptor [6], we applied the descriptor size $\Delta_x(\sigma) = \Delta_y(\sigma) = 2\sigma + 1$, $\Delta_t(\tau) = 2\tau + 1$, where $\sigma = 2$, $\tau = 4$. The PCA was applied to reduce the feature dimensions to 100. For the ESURF descriptor, we used the executable code with default parameter settings [9]: $\Delta_x(\sigma) = \Delta_y(\sigma) = 3\sigma$, $\Delta_t(\tau) = 3\tau$.

For all depth databases, the depth sequences are firstly transformed and stored into gray level videos (depth videos). The skeleton joint positions are also stored for each frame. Then the spatiotemporal features are extracted from the depth videos for each database.

5.2. Evaluation results

The evaluation results are presented in the following, using all four datasets.

5.2.1. On MSRAction3D dataset

MSRAction3D is a commonly used dataset for 3D action recognition. We followed the same settings as [13], where the dataset is divided into 3 subsets, each consisting of 8 actions (see Table 2). Then a cross-subject scheme is used in our evaluation, with half of the subjects for training and the remaining half for testing. The overall accuracy is computed by taking the average over the three subsets. The results of different detectors/descriptors on this dataset are showed in Table 3. One can see that the STIP features have very different accuracies on the same database, ranging from 47.1% to 80.8%, when different detectors and descriptors are used. This observation is very different from the results on color/gray level action videos [3], where the different STIP features
have similar accuracies on the same database. This evaluation indicates the significant difference between 3D depth and color/gray level videos in action recognition.

The highest accuracy is achieved by Harris3D + HOG/HOF feature with a recognition accuracy of 80.8%. This accuracy is comparable to some state-of-the-art approaches, but lower than the highest in the literature by more than 10% (see Table 10 for the state-of-the-art results on MSRAAction3D). Note that in [19] the skeleton joints information was used while in our evaluation of STIP features, only the depth videos are used. One reason that might impact the accuracy is that the interest points cannot be detected for several depth sequences where the lengths of the sequences are quite short.

5.2.2. On MSRDailyActivity3D dataset

The MSRDailyActivity3D dataset contains 16 activities performed by 10 subjects in two scenarios: sitting and standing. Similar to the partition in [13], we divided this dataset into 2 subsets, and evaluate the performance considering two different scenarios, sitting and standing, respectively. We consider the activities in each subset according to the motions: subset 1 (AS1) contains activities without much motion and subset 2 (AS2) with obvious motion. Table 4 shows how we divide the subsets. In our evaluation, we adopt the cross-subject test scheme, using half of the subjects for training and the remaining half for testing. The final results are obtained by averaging accuracies over the subsets.

The evaluation results on MSRDailyActivity3D dataset using different combinations of detectors and descriptors are presented in Table 5. Again, the STIP features achieved very different accuracies. The highest accuracy is obtained by cuboids + HOG/HOF and Hessian + HOG3D, with an accuracy of 70.6%. The result is lower than the reported results, e.g., Oreifej et al. got 80% accuracy with HON4D feature in [22]. The highest accuracy from previous approaches is 85.8% obtained in [19]. In our evaluation, all the combinations of detector/descriptors are above 58%. In the subset with more motion, the performance of STIP is much better (~80%) than the subset with less motion (~50%). This demonstrates that the STIP features can characterize actions with significant motions, but not static actions like sitting. Further, the STIP features cannot represent the human–object interaction. There are several activities in this dataset with similar motion but different objects, e.g., reading and writing, eating and drinking. We also observe that many of the interest points are detected on depth sequences irrelevant to the actions (see Fig. 6). This inspires us to evaluate some refinement schemes for the STIP features (to be shown later).

5.2.3. On UTKinect-action dataset

The evaluation results on the UTKinect-action dataset are showed in Table 6. Note that because many depth sequences in this dataset are of length about 10 frames, which is too short for space–time interest

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Three subsets of actions used for the experiments on MSRAAction3D dataset.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS1</td>
<td>AS2</td>
</tr>
<tr>
<td>Horizontal arm wave</td>
<td>High arm wave</td>
</tr>
<tr>
<td>Hammer</td>
<td>Hand catch</td>
</tr>
<tr>
<td>Forward punch</td>
<td>Draw x</td>
</tr>
<tr>
<td>High throw</td>
<td>Draw tick</td>
</tr>
<tr>
<td>Hand clap</td>
<td>Draw circle</td>
</tr>
<tr>
<td>Bend</td>
<td>Two hand wave</td>
</tr>
<tr>
<td>Tennis serve</td>
<td>Forward kick</td>
</tr>
<tr>
<td>Pickup &amp; throw</td>
<td>Side boxing</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Accuracies of different STIP features on MSRAAction3D dataset. Different detectors and descriptors are combined. Some combinations cannot be realized because of the non-separable executable code.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HOG3D</td>
</tr>
<tr>
<td>Harris3D</td>
<td>76.1%</td>
</tr>
<tr>
<td>Cuboids</td>
<td>77.3%</td>
</tr>
<tr>
<td>Hessian</td>
<td>60.3%</td>
</tr>
</tbody>
</table>

Bold font indicates the highest accuracy.
point detection. Thus a preprocessing is conducted for the depth videos where 10 frames are copied to expand the length of video from both the starting and ending frames.

From the results, the best accuracy is 81%, obtained by Harris3D + HOG3D. This result is lower than the result 90.9% in [14], and the highest accuracy 91.5% in [34]. Note that in [14] and [34] the leave-one-out cross-validation scheme was applied but we use half of the subjects for training and the other half for testing. Fig. 7 shows the confusion matrix of the best STIP feature. Most of the actions are correctly recognized, while the action “carry” has a much lower recognition rate, i.e., 60% of the testing samples are incorrectly classified as “walk”. These two actions are quite similar in the dataset, since “carrying” is performed by a “walking” subject who holds an object. The STIP features might mainly focus on the body motions rather than a relatively small object.

5.2.4. On Cornell activity dataset (CAD-60)

For the CAD-60 dataset, all the depth videos are sampled to 500 frames in our evaluation. All the activity categories (12 desired activities and a random activity) in this dataset are used in our evaluation as in [20]. The same experimental settings are adopted, i.e., three subjects for training, while the remaining for testing.

The evaluation results are shown in Table 7. Among the various features, the Hessian + ESURF gives the highest accuracy 62.5%. From the confusion matrix (Fig. 8), one can see that some of the similar activities on depth sequence are incorrectly recognized, e.g., talkOnCouch and relaxOnCouch, and the random activity in this dataset is also classified as “walking”. These two actions are quite similar in the dataset, since “walking” is performed by a “walking” subject who holds an object. The STIP features might mainly focus on the body motions rather than a relatively small object.

In [20], the precision/recall is reported as the performance measure. [20] reported 71.9%/66.6% on this dataset. In the results, the best accuracy is 81%, obtained by Harris3D + HOG3D. This result is lower than the result 90.9% in [14], and the highest accuracy 91.5% in [34]. Note that in [14] and [34] the leave-one-out cross-validation scheme was applied but we use half of the subjects for training and the other half for testing. Fig. 7 shows the confusion matrix of the best STIP feature. Most of the actions are correctly recognized, while the action “carry” has a much lower recognition rate, i.e., 60% of the testing samples are incorrectly classified as “walk”. These two actions are quite similar in the dataset, since “carrying” is performed by a “walking” subject who holds an object. The STIP features might mainly focus on the body motions rather than a relatively small object.

5.3. Refinements of the STIP features

In the above experiments, various STIP features are evaluated on depth videos with recognition accuracies reported. The best accuracies on each database are comparable to, but lower than some state-of-the-art methods that are developed especially for 3D action analysis.

Table 4

| Subsets of actions used for the experiments on MSRActivity3D dataset |
|-----------------|-----------------|
| A51             | A52             |
| Read book       | Drink           |
| Write on a paper| Eat             |
| Use laptop      | Call cellphone  |
| Use vacuum cleaner | Cheer up   |
| Sit still        | Lay down on sofa|
| Toss paper      | Walk            |
| Play game       | Stand up        |
| Play guitar     | Sit down        |

Table 5

| Accuracies of various STIP features on MSRActivity3D dataset. |
|-----------------|-----------------|
|                  | HOG3D | HOG/HOF | HOG  | HOF  | Cuboids | ESURF |
| Harris3D        | 60.6% | 67.5%   | 63.8%| 59.4%| –       | –     |
| Cuboids         | 68.8% | 76.6%   | 68.1%| 58.1%| 64.4%   | –     |
| Hessian         | 70.0% | 63.8%   | 61.9%| 63.1%| –       | 65.6% |

Table 6

| Accuracies of various STIP features on UTKinect-Action dataset. |
|-----------------|-----------------|
|                  | HOG3D | HOG/HOF | HOG  | HOF  | Cuboids | ESURF |
| Harris3D        | 81.0% | 80.0%   | 66.0%| 69.0%| –       | –     |
| Cuboids         | 65.0% | 65.0%   | 56.0%| 57.0%| 67.0%   | –     |
| Hessian         | 69.0% | 56.0%   | 57.0%| 53.0%| –       | 65.0% |

Bold font indicates the highest accuracy.
some negative impact on action analysis; (2) The overall accuracies on MSRAction3D and MSRArtist3D datasets increase after applying the STIP refinement. On MSRAction3D dataset, the refined accuracy is 80.5%, compared to the original accuracy 78.7%, on MSRArtist3D dataset, the best accuracy is 77.5%, which is higher than the original 70.6%, after the refinement.

5.3.2. STIP feature refinement using RGB images and skeleton joints

We have shown above that in most cases the STIP refinement with the 2D skeleton joint positions can increase the action recognition rates. However, the performance is still highly relied on the interest point detection accuracy. When the interest point detection performs poorly on the depth maps because of the noisy depth data, the skeleton constraints may not help too much. Based on this consideration, we pursue another refinement scheme. The idea is to adopt the interest point detection on RGB videos, i.e., using the STIP locations detected in RGB videos for depth sequences. In other words, the interest point detection is conducted on RGB sequences, and just duplicated to the depth maps. The feature descriptors are still executed on the depth videos.

Experiments are conducted on three datasets except the MSRArtist3D because it does not have the RGB data. We use the same settings as previous. The evaluation results are shown in Table 8. The best STIP feature on each dataset is selected (because separate implementation of ESURF descriptor is not available, we chose the 2nd best STIP feature instead). From the results, one can see that the accuracies are improved significantly after using RGB refinement approach, either the skeleton refinement is applied or not. On MSRArtist3D dataset, the accuracy is increased from 70.6% to 75.6%, on CAD-60 dataset, the accuracy is improved from 56.3% to 68.8%, and on the UTKinect-Action dataset, the accuracy is improved from 81.0% to 85.0%, when using the RGB refinement approach.

For the refinement with skeleton joints, the accuracies can be improved or keep the same on the MSRArtist3D and UTKinect-Action datasets, but reduced on the CAD-60 dataset. The reason could be that the interest points located in the background or scene may help to improve the action recognition accuracies (the CAD-60 dataset contains different actions in different scenes), while the removal of those interest points (constrained by the skeleton joints) can reduce the recognition performance.

The refinement results show that it may not be accurate enough to use the detected locations of interest points on depth sequences directly, because of the noisy depth values.

### Table 8

<table>
<thead>
<tr>
<th>Method</th>
<th>MSRArtist3D</th>
<th>MSRArtist3D</th>
<th>CAD-60</th>
<th>MSRArtist3D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>70.6%</td>
<td>81.0%</td>
<td>56.3%</td>
<td>80.8%</td>
</tr>
<tr>
<td>RGB refined</td>
<td>75.6%</td>
<td>85.0%</td>
<td>68.8%</td>
<td>–</td>
</tr>
<tr>
<td>Skeleton refined</td>
<td>72.5%</td>
<td>84.0%</td>
<td>50.0%</td>
<td>81.7%</td>
</tr>
<tr>
<td>Skeleton &amp; RGB refined</td>
<td>77.5%</td>
<td>85.0%</td>
<td>62.5%</td>
<td>–</td>
</tr>
</tbody>
</table>

Bold font indicates the highest accuracy.

6. Fusing spatiotemporal features and skeleton joints for action recognition

In the above, two approaches have been presented to refine the STIP features. These approaches can be viewed as posing constraints to the interest point locations on depth videos, by using either RGB videos or the skeleton joints. On the other hand, the skeleton joint positions extracted from the depth videos can be used as another feature, representing human posture information. In this section we want to evaluate the performance of combining the STIP features with the skeleton joint feature. This evaluation can tell if the STIP features can complement the skeleton joint features, and if the combination can improve the accuracies significantly. If the accuracies can be improved greatly, it can indicate the usefulness of the STIP features from another aspect.

Specifically, the combination approach has four major steps, which has been presented in a workshop [33]. Firstly, the STIP features are extracted on depth sequences. Then skeleton joint features are computed from the skeleton joint positions. A quantization is performed for the two features respectively to encode the action sequences with histograms. Finally, a feature-level fusion is executed for action recognition using the random forests method [35]. We chose the detector/descriptor combinations which performs the best based on our evaluation presented above. The evaluation of the STIP features in Section 5 is the basis for our fusion approach [33].

We use the histogram of the skeleton joint features proposed in [16] to combine with the best STIP features on each database. Different from [16] where the Naive Bayes classifier was used, we compute the histogram of the joints to combine with the STIP features by the random forests method.

The features from joint locations consist of three parts: (1) current posture: pair-wise joint distances in current posture; (2) motion: joint difference between current posture and the original (in the first frame); and (3) offset: joint differences between current posture and the previous one. A concatenation of the three feature vectors is taken to represent the feature. The PCA method is applied for dimensionality reduction.

To represent each action sequence, we quantize the STIP features and the skeleton joint features, respectively, based on the K-means clustering. The cluster centers are used as the keywords to construct the
histogram bins. These features are used in the next step for feature-level fusion and action classification.

In order to perform the fusion and feature selection of spatiotemporal features and the skeleton joint features, the random forests (RFs) method [35] is used. RFs are usually considered as a classifier using tree predictors in which each tree splits the data depending on the randomly selected features. And there are many nice properties to use the random forests: (1) robustness to noise, (2) efficiency for classification, and (3) the improvement of accuracy by growing multiple trees and vote for the most popular class. Here we use the RFs for fusion of distinct features and action classification together.

The experiments are conducted on the four datasets (MSRAction3D, UTKinect-Action, CAD-60, and MSRDailyActivity3D) while three of them were used in our study in [33]. Our fusion approach can improve the recognition rates to 94.3%, 91.9%, 87.5%, and 80.0%, respectively, on the four databases, which are significantly higher than the STIP feature...
or skeleton (See Table 9.). This result shows that the STIP features can be useful to complement the often-used skeleton features for action recognition.

We also compare the fusion results to other approaches reported in the literature on the four datasets. Table 10 shows all reported results that we can find on the MSRAction3D dataset. Under the same experimental settings, it can be seen that the fusion result of 94.3% accuracy is the second best result among all of the previous methods. Our result is only 0.5% lower than the best result in [36]. On the UTKinect-Action dataset from Table 11, the fusion approach has an accuracy of 91.9% which is higher than the DSTIP + DCSF feature [27], and slightly higher than the HOJ3D feature in [14] (90.9%) and the space–time pose representation in [34] (91.5%). Note that we used the same settings as [27], which is more challenging than the settings in [14] and [34]. On the CAD-60 dataset, the experimental settings are kept the same as [20] and the precision/recall of our fusion method is computed for a direct comparison with other methods, shown in Table 12. Our fusion approach obtained a much higher accuracy than the state-of-the-art results on this dataset. Finally, Table 13 shows the results on the MSRDailyActivity3D dataset, an accuracy 80.0% is obtained using our fusion approach. Slightly different settings are used in our experiment, since the actions are divided into two groups to measure the performance difference between them. Our fusion result is comparable but about 8% lower than the highest accuracy. Note that all the 16 activities are used in our experiment, while in [27], four activities (with less motion) were removed in their experiment.

From the comparison with various approaches, we demonstrate the usefulness of the STIP features for depth-based action recognition, when combined with the skeleton feature.

### 7. Conclusions

We have presented a comprehensive evaluation of the spatio-temporal interest point features for action recognition in 3D. The evaluated STIP features include three spatiotemporal interest point detectors and six descriptors. The combinations of these detectors and descriptors form 14 different features. These STIP features have been evaluated on four different depth action/activity databases. The comparisons to the state-of-the-art methods have shown that the STIP features are still useful for depth-based action recognition.

From the evaluation, we have shown that most of the results are comparable to the current state-of-the-art approaches. However, under...
the bag-of-words framework, the extracted features do not contain the spatial distribution of the interest points in depth maps, this is one reason that limits the performance. We have also shown that the noisy depth data and background have a great impact on interest point detection. Moreover, the interest point detection may not perform well on actions without much motion, resulting in lower accuracies.

The evaluation has shown that different STIP features perform quite differently on depth actions. It discovers that the feature with Harris3D and HOG/HOF performs the best on the MSRAction3D dataset, the cuboid detector with HOG/HOF descriptor performs the best on the MSRDailyActivity3D dataset, while the Harris3D detector combined with HOG3D descriptor is the best on UTKinect-Action dataset. On the CAD-60 dataset, the Hessian detector with ESURF descriptor gives the highest accuracy.

Two interest point refinement schemes have been presented for the STIP features, based on constraining the STIP features using skeleton joint positions and/or the detection in RGB videos. We have shown that the STIP features can be refined to achieve better performance in most cases. We have also proposed a fusion scheme to combine the best STIP features with the skeleton joint features in each database. Significant improvements of the recognition accuracies have been achieved on all four databases. Overall, we have explored the STIP features for 3D action recognition from different aspects.

Acknowledgment

The authors thank the anonymous reviewers for their detailed comments to improve the paper.

Table 11
Comparisons of different methods on UTKinect-Action dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSTIP + DCSF [27]</td>
<td>85.8%</td>
</tr>
<tr>
<td>HOJ3D [14]</td>
<td>90.0%</td>
</tr>
<tr>
<td>Space-time pose representation [34]</td>
<td>91.5%</td>
</tr>
<tr>
<td>STIP + skeleton</td>
<td>91.9%</td>
</tr>
</tbody>
</table>

Table 12
Comparisons of different methods on CAD-60 dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision/recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>J. Sung et al. [20]</td>
<td>67.9%/55.5%</td>
</tr>
<tr>
<td>X. Yang et al. [16]</td>
<td>71.9%/66.6%</td>
</tr>
<tr>
<td>Koppula et al. [23]</td>
<td>80.8%/71.4%</td>
</tr>
<tr>
<td>STIP + skeleton</td>
<td>93.2%/84.6%</td>
</tr>
</tbody>
</table>

Table 13
Comparisons of different methods on MSRDailyActivity3D dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBNN + parts + time [37]</td>
<td>70.0%</td>
</tr>
<tr>
<td>Local HON4D [22]</td>
<td>80.0%</td>
</tr>
<tr>
<td>DCSF [27]</td>
<td>83.6%</td>
</tr>
<tr>
<td>RGGF + fusion [38]</td>
<td>85.6%</td>
</tr>
<tr>
<td>Actionlet [19]</td>
<td>85.8%</td>
</tr>
<tr>
<td>DCSF + joint [27]</td>
<td>88.2%</td>
</tr>
<tr>
<td>STIP + skeleton</td>
<td>80.0%</td>
</tr>
</tbody>
</table>
References


