Image and Vision Computing 32 (2014) 453–464

Contents lists available at ScienceDirect

Image and Vision Computing

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

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

ARTICLE INFO

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

ABSTRACT

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.

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

[☆] This paper has been recommended for acceptance by Ioannis Patras .
* Corresponding author.
E-mail address: guodong.guo@mail.wvu.edu (G. Guo).
[1] The first and second authors have equal contributions.

http://dx.doi.org/10.1016/j.imavis.2014.04.005
0262-8856/© 2014 Elsevier B.V. All rights reserved.

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 zy, 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 (HOJ3D) 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 to the horizontal plane, 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 HMM 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 Naive-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 $F_i(\mathrm{t}) = (x_i(\mathrm{t}), y_i(\mathrm{t}), z_i(\mathrm{t}))$ at frame t, the pairwise relative position features are extracted for joint i as: $p_i = \{p_{ij} | i \neq j\} = \{p_i - p_j | i \neq j\}$. 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: $o_{xyz} = \delta\left(\sum_{q \in bin_{xyzt}} I_q\right)$, where $I_q = 1$ if the point cloud has a point in the location q and $I_q = 0$ otherwise. $\delta(\cdot)$ is a sigmoid normalization function: $\delta(x) = \frac{1}{1+e^{-\beta x}}$. 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 polychoron 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 $L_1^z = [z_1^l, z_1^u]$, $L_2^z = [z_2^l, z_2^u]$, ..., $L_M^z = [z_M^l, z_M^u]$, with lower and upper boundaries denoted as z_M^l and z_M^u 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 HOGHOF 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.