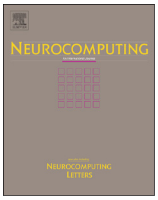
Neurocomputing 175 (2016) 747–758

Contents lists available at ScienceDirect

Neurocomputing

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



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

Mengyuan Liu, Hong Liu*

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

ARTICLE INFO

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

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

ABSTRACT

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

1. Introduction

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

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

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

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

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

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

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

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

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

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

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

2. Overview of the proposed framework

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

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

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

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