Pattern Recognition Letters 36 (2014) 221–227

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

On the improvement of human action recognition from depth map sequences using Space–Time Occupancy Patterns

Antonio W. Vieira[a,b,*], Erickson R. Nascimento[a], Gabriel L. Oliveira[a], Zicheng Liu[c], Mario F.M. Campos[a]

[a] DCC – Universidade Federal de Minas Gerais, Belo Horizonte 31270-010, Brazil
[b] CCET – Unimontes, Montes Claros, Brazil
[c] Microsoft Research, Redmond, USA

ARTICLE INFO

Article history:
Available online 25 July 2013

Communicated by Luis Gomez Deniz

Keywords:
Pattern recognition
Human action
Depth map

ABSTRACT

We present a new visual representation for 3D action recognition from sequences of depth maps. In this new representation, space and time axes are divided into multiple segments to define a 4D grid for each depth map sequences. Each cell in the grid is associated with an occupancy value which is a function of the number of space–time points falling into this cell. The occupancy values of all the cells form a high dimensional feature vector, called Space–Time Occupancy Pattern (STOP). We then perform dimensionality reduction to obtain lower-dimensional feature vectors. The advantage of STOP is that it preserves spatial and temporal contextual information between space and time cells while being flexible enough to accommodate intra-action variations. Furthermore, we combine depth maps with skeletons in order to obtain view invariance and present an automatic segmentation and time alignment method for on-line recognition of depth sequences. Our visual representation is validated with experiments on a public 3D human action dataset.

1. Introduction

Human action recognition has a wide range of applications including senior home monitoring, video surveillance, video indexing and search, human robot interaction, and entertainment, to name a few. So far, most of the work has been focused on using 2D video sequences as input due to the ubiquity of conventional video cameras. Recently, low cost, commercially available, RGB-D sensors, such as the Microsoft Kinect, has rapidly driven the popularity of the depth sensors. The real time depth maps captured by this sensor is fostering the development of enhanced methodologies for human action recognition and interaction.

State-of-the-art algorithms for action recognition use silhouettes, Space–Time Interest Point (STIP) and skeletons. Skeletons can be obtained from motion capture systems using body joint markers or directly tracked from depth maps. However, tracking of body joints from depth maps is not a completely solved problem. For example, Li et al. (2010) reported that the joint positions returned by the XBOX Kinect skeleton tracker are quite noisy. As also reported by those authors, for 3D action recognition, the performance of their method using joint position was worse than using bag of 3D points. In addition, most of the current real-time depth cameras can only produce coarse and noisy depth maps, making skeleton tracking even more challenging.

Recently, some works have been published to address action recognition from depth map sequences. The work in Li et al. (2010) uses the silhouettes projected onto the coordinate planes for the depth maps in any given frame, and sample a small set of 3D points, which are the interest points. For each interest point, they use the 3D coordinates as feature descriptor. The bag (collection) of these points constitutes the visual representation of the frame. The dissimilarity between two depth maps is computed by the Hausdorff distance between the two sets of interest points. One limitation of this approach is that the spatial context information between interest points is lost. Furthermore, due to noise and occlusions in the depth maps, the silhouettes viewed from the side and from the top may not be very reliable. This makes it very difficult to robustly sample the interest points given the geometry and motion variations between different persons. This is probably why they reported low recognition accuracy for the cross-subject test which is much worse than the accuracy attained with the other two non-cross-subject tests presented in their paper.

In Yang et al. (2012), propose a new type of feature based on position differences of joints, called *EigenJoints*, which combine action information including static posture, motion, and offset. They employ the Naïve–Bayes-Nearest-Neighbor classifier for multi-class action classification and also explore the number of frames that are needed to classify an action in a depth sequence. In

* Corresponding author at: DCC – Universidade Federal de Minas Gerais, Belo Horizonte 31270-010, Brazil. Tel.: +55 (38) 9966 0068; fax: +55 (31) 3409 5858.

E-mail addresses: awilson@dcc.ufmg.br (A.W. Vieira), erickson@dcc.ufmg.br (E.R. Nascimento), gabrielleivas@gmail.com (G.L. Oliveira), zliu@microsoft.com (Z. Liu), mario@dcc.ufmg.br (M.F.M. Campos).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.07.011

Yang et al. (2012), project depth maps onto three orthogonal planes and accumulate global activities across entire video sequences to generate the Depth Motion Maps (DMM). Histograms of Oriented Gradients (HOG) are then computed from DMM as the representation of an entire action video. These off-line methods considers sequences segmented out in start and end frames for classifying the entire sequence.

Our approach presents a new feature called Space–Time Occupancy Pattern (STOP) where depth sequences is represented in a 4D space–time grid and uses a saturation scheme to enhance the roles of the sparse cells which typically consist of points on the silhouettes or moving parts of the body. These cells contain important information for action recognition and we show that the feature vectors obtained by using this scheme perform much better than the original histogram vectors without saturation. For comparison purpose, we present an off-line classification scheme using our STOP features for classifying entire sequences. In addition, we propose an on-line classification scheme, where an action graph based system is used to learn a statistical model for each action class and use a state machine to segment long depth sequences using a neutral pose classifier.

We evaluated our technique performing several experiments in the public MSR Action3D Dataset (Li et al., 2010). The experiments show that our technique achieves recognition accuracy comparable to those obtained by existing off-line methods. Furthermore, we present results for unsegmented depth sequences to show performance of our system for on-line classification.

The remainder of the paper is organized as follows: Section 2 briefly reviews related works. Section 3 describes our STOP features and classification scheme. Experimental results are shown in Section 4 and, finally, Section 5 presents our conclusions and future work directions.

2. Related work

The action recognition methods can be classified either as global or as local methods. The methods in the first class use global features such as silhouettes (Lv et al., 2007; Li et al., 2008) and space–time volume information (Gorelick et al., 2007; Yilmaz and Shah, 2005). The methods in the second class use local features for which a set of interest points are extracted from a video and a feature descriptor is computed for each interest point. Those locally extracted features are used to characterize actions for recognition (Dollar et al., 2005; Sun et al., 2009). We refer the reader to the excellent survey in Weinland et al. (2011) on action representation, segmentation and recognition.

The amount of work on action recognition from 3D data has been quite limited when compared to those that use 2D data, due to the difficulty in acquisition of 3D data acquisition.

One way to obtain 3D data is by using marker-based motion capture systems (MoCap). Such systems capture the 3D positions of markers which are positioned as close as possible to the joints of the body of a performer. Joint positions captured by such systems are, in general, quite accurate. One dataset obtained with a MoCap system can be downloaded from http://mocap.cs.cmu.edu (2012). Lv and Nevatia (2006) used this dataset for action recognition experiments where joint positions were used as the basic features, while Vieira et al. (2012) showed that joint distance matrices provide invariant features for classifying MoCap data. Han et al. (2010) developed a technique to learn a low-dimensional subspace from the high dimensional space of joint positions, and performed action recognition in the learned low-dimensional space.

A second way to obtain 3D data is to use multiple 2D video streams to reconstruct 3D information. Huang et al. (2005) developed a system that uses multiple omni-directional cameras to capture a scene and to reconstruct the 3D volumetric data with a visual-hull based technique (Laurentini, 1994). They also proposed a 3D shape context representation for action recognition.

Another visual-hull based technique used for 3D volumetric reconstruction is proposed by Weinland (2006). They used motion history volumes for action recognition. Similar to Weinland (2006), Gu et al. (2010) developed a system to create volumetric data from multiple views of a scene. They also recovered the joint positions which were used as features for action and gait recognition.

The third way to obtain 3D data is using depth sensors. One of the existing types of depth sensors is based on the time-of-flight principle (Iddan, 2001). Such sensors are called *time-of-flight* cameras. These cameras have been used in several recognition systems such as hand gesture recognition (Liu et al., 2004) and template matching of 3D articulated hands (Breuer et al., 2007). Another type of depth sensor is based on structured light patterns. A large number of systems that use visible structured light patterns exists, of which Malassiotis et al. (2001) is a typical example.

Visible light patterns have a drawback that much of the visual content of a scene is in the visible spectrum, and the projection becomes quite invasive and disrupts several applications. Infra-red structured lights Ypsilos et al. (2004) have thus become an attractive alternative. Recently, Microsoft released a depth camera, called Kinect, which is based on the projection of infrared structured pattern. Li et al. (2010) developed a technique for action recognition from depth maps captured by a depth camera similar to Kinect, and produced a dataset with various people performing different actions. State-of-the-art results on this dataset are presented in Yang et al. (2012), Yang et al., 2012. In this work we also use the same dataset, both for validation and for comparative purposes.

3. Space–Time Occupancy Patterns

Our visual representation is in part inspired on the well known occupancy grid approach, which is commonly used for robot navigation (Elfes, 1989). A 2D plane or 3D space is divided into a grid where to each cell is assigned a probability indicating the certainty of its occupation. To construct our visual representation, we consider a sequence (in time) of depth data acquired from a person performing an action as a set

$$A = \{(x_i, y_i, z_i, t_i), i = 1, \dots, N\}, \tag{1}$$

into a space–time box B, where the fourth dimension is the frame index t_i, which indicates the time of acquisition. This space–time box is then partitioned into a four-dimensional grid with m 4D cells. Let x, y, z, and t denote the four axes, $B \subset \mathbb{R}^4$ denotes the space–time box and denote by n_x, n_y, n_z, and n_t the number of segments divided uniformly along x, y, z, and t axes, respectively. Then, B is partitioned into a grid with m 4D cells. We use c_i to denote the ith cell. The set of cells is called a *partition*, denoted as $C = \{c_1, \dots, c_m\}$. For each cell c_i, we denote by A_i its intersection with the set of four-dimensional points A, that is, $A_i = A \cap c_i$. The occupancy value of c_i is defined as

$$P(c_i) = \begin{cases} 1, & \text{if} \quad |A_i| \geqslant p \\ \frac{|A_i|}{p}, & \text{otherwise} \end{cases}, \tag{2}$$

where p is a predefined saturation parameter, empirically selected to maximize recognition accuracy. For each cell that contains p or more points, its occupancy value is set to the maximum value of 1.0. The reason for this saturation scheme is because the number of points contained in a nonempty cell may be as small as 1, and as large as several thousands. For a typical sequence, the majority of the nonempty cells contains only a few hundred points. On the one hand, if the histogram is used directly without saturation, the cells that contain a small number of points would not play any significant role in the classification step since their values would be