Pattern Recognition 61 (2017) 222–233

Contents lists available at ScienceDirect

Pattern Recognition

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

Motion segment decomposition of RGB-D sequences for human behavior understanding

Maxime Devanne [a,b,*], Stefano Berretti [b], Pietro Pala [b], Hazem Wannous [a], Mohamed Daoudi [a], Alberto Del Bimbo [b]

[a] Télécom Lille, Univ. Lille, CNRS, UMR 9189 - CRIStAL, F-59000 Lille, France
[b] MICC/University of Florence, Florence, Italy

ARTICLE INFO

Article history:
Received 12 February 2016
Received in revised form
23 June 2016
Accepted 27 July 2016
Available online 28 July 2016

Keywords:
3D human behavior understanding
Temporal modeling
Shape space analysis
Online activity detection

ABSTRACT

In this paper, we propose a framework for analyzing and understanding human behavior from depth videos. The proposed solution first employs shape analysis of the human pose across time to decompose the full motion into short temporal segments representing elementary motions. Then, each segment is characterized by human motion and depth appearance around hand joints to describe the change in pose of the body and the interaction with objects. Finally, the sequence of temporal segments is modeled through a Dynamic Naive Bayes classifier, which captures the dynamics of elementary motions characterizing human behavior. Experiments on four challenging datasets evaluate the potential of the proposed approach in different contexts, including gesture or activity recognition and online activity detection. Competitive results in comparison with state-of-the-art methods are reported.

1. Introduction

Visual recognition and understanding of human activity and behavior represent a task of interest for many multimedia applications, including entertainment, medicine, sport, video surveillance, human–machine interfaces and active assisted living. This wide spectrum of potential applications encouraged computer vision community to address the issue of human behavior understanding from 2D videos taken from standard RGB cameras [1–5]. However, most of these methods suffer from some limitations, like the sensitivity to color and illumination changes, background clutter and occlusions. Since the recent release of RGB-D sensors, new opportunities have emerged in the field of human motion analysis and understanding. Hence, many research groups investigated data provided by such cameras in order to benefit from some advantages compared to RGB cameras [6–10]. Indeed, depth data allows a better understanding of the 3D structure of the scene and thus makes background subtraction and people detection easier. In addition, the technology behind such depth sensors provides robustness to light variations as well as the capability to work in complete darkness. Finally, the combination of such depth sensors and powerful pattern recognition algorithms [11] enables the representation of human pose at each frame as a set of 3D joints. In the past decades, human motion analysis from 3D data provided by motion capture systems has been widely investigated [12–14]. While these systems are very accurate, they present some disadvantages. First, the cost of such technology may limit its use. Second, it implies that the subject wears some physical markers so as to estimate the 3D pose. As a result, this technology is not convenient for the general public. All these considerations motivated us to focus our study of human behavior on RGB-D data. However, this task still faces some major challenges due to the temporal variability and complexity of human actions and the large number of motion combinations that can characterize the human behavior. Motion analysis is further complicated by the fact that it should be invariant to geometric transformations, such as translation, rotation and global scaling of the scene. In addition, human behavior often involves interaction and manipulation of objects. While such information about the context may help the understanding of what the human is doing, it also involves possible occlusions of parts of the human body, resulting in missing or noisy data.

In order to face these challenges, we propose in this paper to locally investigate the sequence by detecting short temporal segments representing elementary motion, called *Motion Segments* (MS). Then, for each MS, we analyze human motion and depth appearance around human hands to characterize the interaction with objects. This provides a deeper analysis of the human behavior and allows the recognition of human *gestures*, *actions* and *activities*. In particular, in this paper, *gestures* indicate simple movements performed with only one part of the body, *actions* represent a combination of gestures with different parts of the

* Corresponding author at: Telecom Lille, Rue Gluglielmo Marconi, 59650 Villeneuve d'Ascq, France.
E-mail address: maxime.devanne@telecom-lille.fr (M. Devanne).

http://dx.doi.org/10.1016/j.patcog.2016.07.041
0031-3203/

body, and *activities* refer to more complex motion patterns possibly involving interaction with objects. The proposed solution can be adapted to realistic scenarios, where several actions or activities are performed subsequently in a continuous sequence. In that case, the sequence should be processed *online* in order to detect the starting and ending time of actions or activities. That is, the proposed approach can operate on the data stream directly, without assuming the availability of a segmentation module that identifies the first and last frame of each action/activity.

1.1. Previous work

In recent years, recognition and understanding of human behavior by analyzing depth data have attracted the interest of several research groups [15–18]. While some methods focus on the analysis of human motion in order to recognize human *gestures* or *actions*, other approaches try to model more complex behaviors (*activities*) including object interaction. These solutions focus on the analysis of short sequences, where one single behavior is performed along the sequence. However, additional challenges appear when several different behaviors are executed one after another over a long sequence. In order to face these challenges, methods based on *online detection* have been proposed. Such methods can recognize behavior before the end of their execution by analyzing short parts of the observed sequence. Thus, these methods are able to recognize multiple behaviors within a long sequence, which may not be the case for methods analyzing the entire sequence directly. Existing methods for human behavior recognition using depth data are shortly reviewed below.

Methods analyzing human motion for the task of *gesture* / *action* recognition from RGB-D sensors can be grouped into three categories: *skeleton*-based, *depth map*-based and *hybrid* approaches. Skeleton based approaches have become popular thanks to the work of Shotton et al. [11]. This describes a real-time method to accurately predict the 3D positions of body joints in individual depth maps, without using any temporal information. In [19], Yang and Tian performed human action recognition by extracting three features for each joint, based on pair-wise differences of joint positions (initial, previous and current frames). PCA is then used to obtain a compact *EigenJoints* representation of each frame and a naïve-Bayes nearest-neighbor classifier is used for multi-class action classification. Similar features are used by Luo et al. [20], but pairwise differences are computed only in the current frame and with respect to only one reference joint (the hip joint). To better represent these features, they propose a dictionary learning method based on group sparsity and geometry constraints. The classification of sequences is performed using SVM. Zanfir et al. [15] propose the Moving Pose feature, capturing for each frame the human pose information as well as the speed and acceleration of body joints within a short temporal window. A modified *kNN* classifier is employed to perform action recognition. Hongzhao et al. [21] introduce a part-based feature vector to identify the most relevant body parts in each action sequence. Other approaches use differential geometry to represent skeleton data. In [22], Vemulapalli and Chellappa represented each skeleton as one element on the Lie-group, and the sequence corresponds to a curve on this manifold. In [23], Slama et al. express the time series of skeletons as one point on a Grassmann manifold, where the classification is performed benefiting from Riemannian geometry of this manifold. In [24], Anirudh et al. regard actions as trajectories on a Riemannian manifold, and analysis of such trajectories using Transport Square-Root Velocity Function is employed for action recognition.

Methods based on depth maps extract meaningful descriptors from the entire set of points of depth images. In [25], Yang et al. described the action dynamics using Depth Motion Maps, which highlight areas where some motion takes place. Other methods, such as Spatio-Temporal Occupancy Pattern [26], Random Occupancy Pattern [27] and Depth Cuboid Similarity Feature [16], propose to work on the 4D space divided into spatio-temporal boxes to extract features representing the depth appearance in each box. Such features are extracted from Spatio-Temporal Interest Points. A similar method is proposed by Rahmani et al. [28], where keypoints are detected and the point cloud is described within a volume using the Histogram of Principal Components. In [29], Oreifej and Liu proposed a method to quantize the 4D space using vertices of a polychoron, and then model the distribution of the normal vectors for each cell. The idea of using surface normals to describe both local motion and shape information characterizing human action is also used by Yang and Tian [30]. Althloothi et al. [31] represent 3D shape features based on spherical harmonics representation and 3D motion features using kinematic structure from skeleton. Both features are then merged using a multi-kernel learning method. A depth feature to describe shape geometry and motion, called Range-Sample, is proposed by Lu and Tang [32].

Analyzing human motion, however, may not be sufficient to understand more complex behaviors involving human interaction with the environment (i.e., what we call *activities*). Hybrid solutions are often proposed, which use depth maps for modeling scene objects and body skeleton for modeling human motion. For example, Wang et al. [33] used Local Occupancy Patterns to represent the observed depth values in correspondence to skeleton joints. Other methods propose to describe and model spatio-temporal interaction between human and objects characterizing the activities, using Markov Random Field [17]. A graphical model is also employed by Wei et al. [34] to hierarchically define activities as combination of sub-events including description of the human pose, the object and interaction between them. Yu and Liu [35] propose to capture meaningful skeleton and depth features using a middle level representation called *orderlet*.

Some of the works reviewed above have also *online* action recognition capabilities, as they compute their features within a short sliding window along the sequence [35]. This challenge has recently been investigated for continuous depth sequences, where several actions or activities are performed successively. For example, Huang et al. [18] proposed and applied the Sequential Max-Margin Event Detector algorithm on long sequences comprising many actions in order to perform online detection by successively discarding not corresponding action classes.

1.2. Overview of our approach

Human behavior is naturally characterized by the change of the human body across time. Thanks to depth sensors, we are able to capture skeleton data containing the 3D position of different parts of the body. The skeleton and its changes across time provide valuable information. However, understanding the human behavior is still a difficult task due to the complexity of human motion and spatial/temporal variations in the way gestures, actions, or activities are performed. These challenges motivated us to analyze locally the motion sequences. First, we represent the skeleton of each frame by a 3D curve describing human pose. These curves are then interpreted in a Riemannian manifold, which defines a *shape space* where shapes of the curves can be modeled and compared using elastic registration and matching. Such shape analysis allows the identification and grouping of the human poses. As a result, a motion sequence is temporally segmented into a set of successive sub-sequences of elementary motions, called *Motion Segments* (MS). A MS is thus characterized by a sequence of skeletons, each of which is modeled as a multi-dimensional vector by concatenating the three-dimensional coordinates of its joints. Then, the trajectory described by this vector in