Pattern Recognition ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Pattern Recognition

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

3D skeleton-based human action classification: A survey

Liliana Lo Presti *, Marco La Cascia

V. le delle Scienze, Ed. 6, University of Palermo, 90128 Palermo, Italy

ARTICLE INFO

Article history:
Received 1 April 2015
Received in revised form
27 October 2015
Accepted 24 November 2015

Keywords:
Action recognition
Skeleton
Body joint
Body pose representation
Action classification

ABSTRACT

In recent years, there has been a proliferation of works on human action classification from depth sequences. These works generally present methods and/or feature representations for the classification of actions from sequences of 3D locations of human body joints and/or other sources of data, such as depth maps and RGB videos.

This survey highlights motivations and challenges of this very recent research area by presenting technologies and approaches for 3D skeleton-based action classification. The work focuses on aspects such as data pre-processing, publicly available benchmarks and commonly used accuracy measurements. Furthermore, this survey introduces a categorization of the most recent works in 3D skeleton-based action classification according to the adopted feature representation.

This paper aims at being a starting point for practitioners who wish to approach the study of 3D action classification and gather insights on the main challenges to solve in this emerging field.

1. Introduction

In several application domains, such as surveillance [1–4], human–computer interaction [5], assistive technologies [6], sign language [7–9], computational behavioral science [10,11] and consumer behavior analysis [12], the focus is the detection, recognition and analysis of human actions and behaviors. These applications have motivated a large part of the computer vision community to conduct research on action recognition and modeling. Therefore, there is a vast literature, including interesting surveys [13–19], on human body pose estimation [20–24] and activity/action classification [25–30].

As often happens in computer vision, there are psychological studies that motivate current approaches. In the case of human body pose estimation for action classification, Johansson's moving light-spots experiment [31] for motion perception[1] is certainly the most notable. This experiment was conducted in the 1970s to study 3D human motion perception from 2D patterns. The study has the goal of analyzing the visual information from typical motion patterns when some pictorial form aspect of the patterns is known. To these purposes, several bright spots distributed on the human body against a homogeneous, contrasting background are used in the experiment (see Fig. 1). The experiment demonstrates that the number of light-spots and their distribution on the human body may affect motion perception. In particular, an increasing number of light-spots may decrease ambiguity in motion understanding.

Johansson's study demonstrates that human vision not only detects motion directions but can also detect different types of limb motion patterns, including recognition of the activity and of the velocity of the different motion patterns. As reported in [31], "The geometric structures of body motion patterns in man […] are determined by the construction of their skeletons. […] From a mechanical point of view, the joints of the human body are end points of bones with constant length […] ".

This study has inspired most of the literature about human body pose estimation and action recognition [32–34] as, by knowing the position of multiple body parts, we want the machine to learn to discriminate among action-classes.

In particular, works on human body pose estimation try to estimate the configuration of the body (pose) typically from a single, monocular, image [35]. Part detectors [36,37] and/or pictorial structure (PS) models [38–40] are used to model the appearance of body parts and infer the pose based on constraints among body parts; in general, such constraints are meant to represent the actual human body articulations.

The major difficulty in body-pose estimation is that human body is capable of an enormous range of poses, which are also difficult to simulate or to account for. Technologies such as motion capture (Mo-Cap) have been used to collect accurate data and the corresponding ground-truth. Due to the difficulty in reliably estimating the body pose, several approaches have tried to use holistic representation of the body pose. Methods such as [27,41,42] have proved to be successful in performing recognition of simple

* Corresponding author. Tel.: +39 09123899526.
E-mail address: liliana.lopresti@unipa.it (L. Lo Presti).

[1] See for example https://www.youtube.com/watch?v=1F5ICP9SYLU.

http://dx.doi.org/10.1016/j.patcog.2015.11.019
0031-3203/

actions while skipping the human body pose inference step and adopting features that are correlated with the body pose. Emerging trends are action recognition "in the wild" [43–48], and action "localization" [49–53].

Technologies are evolving fast, and the very recent wide diffusion of cheap depth camera, and the seminal work by Shotton et al. [56] for estimating the joint locations of a human body from depth map have provided new stimulus to the research in action recognition given the locations of body joints. Depth map proved to be extremely useful in providing data for an easy and fast human body estimation. As first introduced in [31], the computer vision community defines a *skeleton* as a schematic model of the locations of torso, head and limbs of the human body. Parameters and motion of such a skeleton can be used as a representation of gestures/actions and, therefore, the human body pose is defined by means of the relative locations of the joints in the skeleton. Fig. 2 shows a graphical representation of possible skeletons (estimated by the methods [56,57]) where the red dots represent the estimated human body joints.

Gaming applications and human–computer interfaces are benefiting of the introduction of these new technologies. In the very recent years, we have assisted to a proliferation of works introducing novel body pose representations for *3D action classification* given depth maps and/or skeleton sequences [58–60]. These works have shown how, even with a reliable skeleton estimation, 3D skeleton-based action classification is not that simple as it may appear. In this sense, "one of the biggest challenges of using posed-based features is that semantically similar motions may not necessarily be numerically similar" [61]. Moreover, motion is ambiguous and action classes can share movements [62]. Most of the works in 3D action recognition attempt to introduce novel body pose representations for action recognition from 3D skeletal data [55,34,58] with the goal of capturing the correlation among different body joints across time. Other works [63,64] attempt to mine the most discriminative joints or group of joints for each class. Promising works [65,62] deal with the dynamics regulating the 3D trajectories of the body joints. Across all these works, various different classification frameworks have been applied to classify the actions.

In contrast to very recent surveys such as [66,67] which focus on the use of depths cameras in several application domains or present a review of the literature on activity recognition in depth maps, in this survey we focus mostly on action classification from skeletal data. To this purpose, we present an overview of the technologies used for collecting depth maps (see Section 2) and review the most used methods at the state of the art for estimating skeletons/body pose in Section 3.

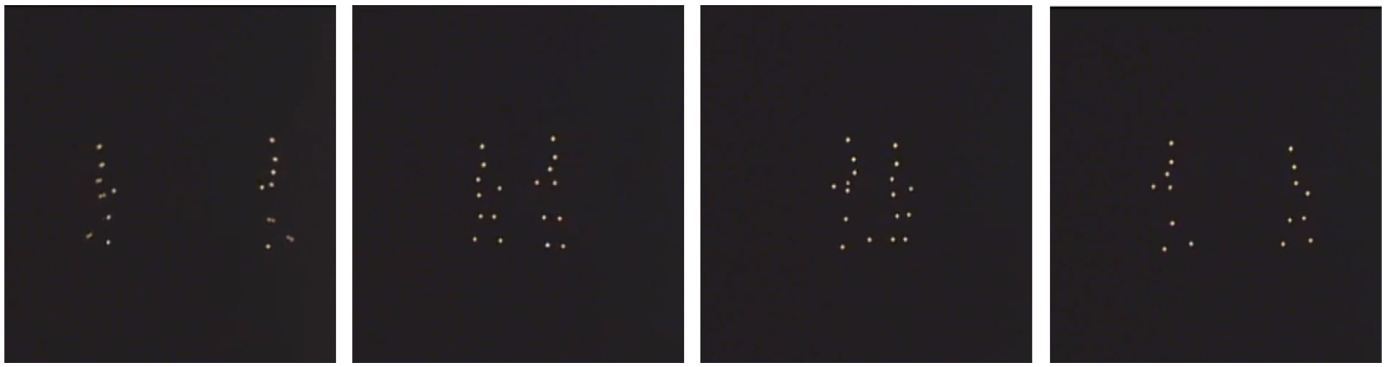

Fig. 1. Four frames of the video (see footnote 1) showing Johansson's moving light-spots experiment: two persons cross each other.

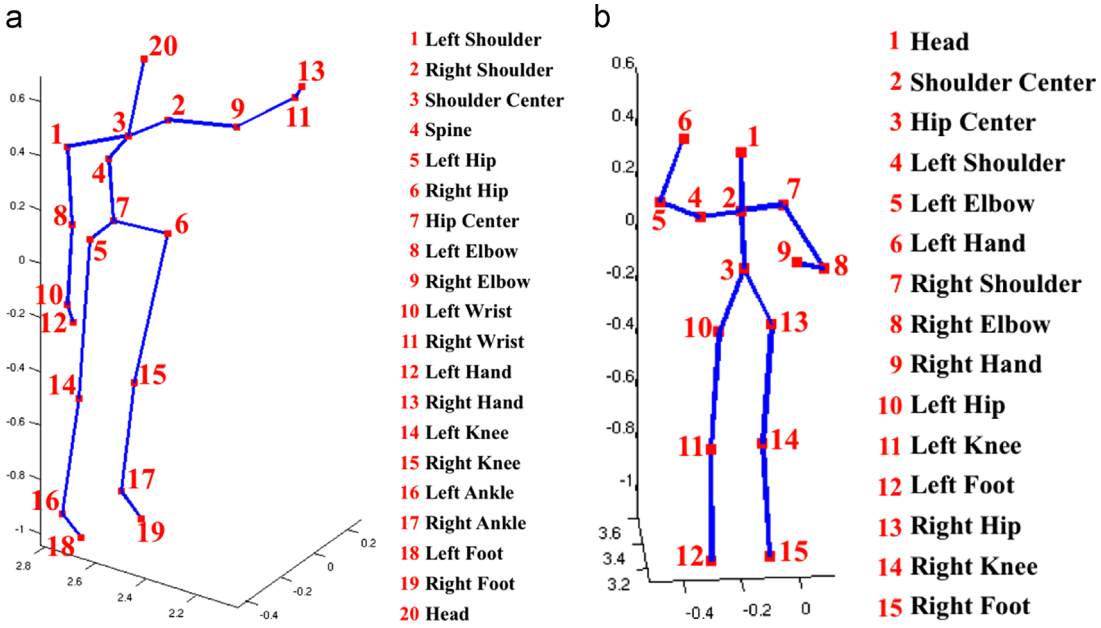

Fig. 2. Graphical representation of skeletal data with 20 and 15 joints. (a) Skeleton of 20 joints (MSRA-3D dataset [54]) and (b) skeleton of 15 joints (UCF Kinect dataset [55]). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Please cite this article as: L. Lo Presti, M. La Cascia, 3D skeleton-based human action classification: A survey, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.11.019