Robotics and Autonomous Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

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

A multi-viewpoint feature-based re-identification system driven by skeleton keypoints

Stefano Ghidoni *, Matteo Munaro

Department of Information Engineering, University of Padova, via Gradenigo, 6/B – 35131 Padova, Italy

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
People re-identification
People tracking
Body pose estimation
Camera networks
Multi-view skeletal tracker

ABSTRACT

Thanks to the increasing popularity of 3D sensors, robotic vision has experienced huge improvements in a wide range of applications and systems in the last years. Besides the many benefits, this migration caused some incompatibilities with those systems that cannot be based on range sensors, like intelligent video surveillance systems, since the two kinds of sensor data lead to different representations of people and objects. This work goes in the direction of bridging the gap, and presents a novel re-identification system that takes advantage of multiple video flows in order to enhance the performance of a skeletal tracking algorithm, which is in turn exploited for driving the re-identification. A new, geometry-based method for joining together the detections provided by the skeletal tracker from multiple video flows is introduced, which is capable of dealing with many people in the scene, coping with the errors introduced in each view by the skeletal tracker. Such method has a high degree of generality, and can be applied to any kind of body pose estimation algorithm. The system was tested on a public dataset for video surveillance applications, demonstrating the improvements achieved by the multi-viewpoint approach in the accuracy of both body pose estimation and re-identification. The proposed approach was also compared with a skeletal tracking system working on 3D data: the comparison assessed the good performance level of the multi-viewpoint approach. This means that the lack of the rich information provided by 3D sensors can be compensated by the availability of more than one viewpoint.

1. Introduction

People re-identification is a topic that has been widely addressed in the literature, as it is a crucial capability in several fields, including intelligent video surveillance, service robotics, assistive robotics and many others. While people re-identification is a widely addressed topic in the field of image processing, techniques based on the analysis of 3D data became popular only recently. This was caused by the introduction of low-cost, high-resolution RGB-D (RGB-Depth) sensors into the market, which pushed mobile robotics towards 3D vision and gave strong impulse to the development of the Point Cloud Library (PCL)[1] [1]. First released in 2011, four years later it is considered the *de facto* standard for the processing, filtering and storage of 3D data, with the plus of being open source, and it is used throughout the world.

The introduction of RGB-D data had a strong impact on many applications connected to mobile robotics. As an example, consider people tracking systems: moving from RGB to RGB-D sensing provides, on one hand, enhanced accuracy in the metric measurements (e.g. target location and dimensions). On the other hand, this offers new cues that can be observed and exploited for tracking purposes, like the body shape. This also had a positive impact on the algorithms for skeletal tracking (i.e., the detection of the body pose): those based on 3D data provide reliable results, while the 2D counterparts still show some difficulties, even though great improvements have been achieved [2].

The picture outlined so far suggests that 3D sensing is the way to move robot vision one step forward; however, there are some applications and environments that are more suitable for 2D vision. The most important limiting factor is the range of the mentioned low-cost 3D sensors, that can be (in the case of the Kinect 2) up to 8–9 m in the best working conditions [3]. This means that robot navigation and people and object detection cannot be based on 3D sensing in large industrial environments, or outdoors (where the working range is reduced) unless more sophisticated and expensive sensors are employed, as it is the case of the Velodyne sensor [4]. Likewise, intelligent video surveillance applications usually deal with events that occur at a rather long distance, and are rarely based on RGB-D sensors.

In summary, the availability of low-cost RGB-D sensors created two different research lines that tackle similar problems based on different sensory information. While this gave new impulse to the research activity and generally increased the performance level of

* Corresponding author.
E-mail address: ghidoni@dei.unipd.it (S. Ghidoni).

[1] Available at: www.pointclouds.org.

http://dx.doi.org/10.1016/j.robot.2016.10.006
0921-8890/

robotic perception, an important drawback was also introduced: a more difficult interoperability among systems based on different sensors. As an example, consider a service robot moving in an environment where an intelligent video surveillance system making use of a camera network is also operating. If the robot is running a people tracking algorithm based on RGB-D data, it will make use of models that rely on 3D information: thus, a comparison between its tracks and those generated by the video surveillance system is difficult to perform.

This paper goes in the direction of filling the gap: a novel re-identification system for camera networks is presented, based on the analysis of RGB data only, that is, 2D images, without relying on 3D information. The lack of depth data is counterbalanced by the observation of the scene from multiple viewpoints, that provides additional information, exploited to improve the performance of the body pose estimation. The comparison between the multi-viewpoint 2D approach and an open source single-viewpoint 3D skeletal tracker, presented in Section 4, demonstrates that relying on 3D data is not always the best choice.

The system presented in this paper aims at generating a body model suitable for re-identification which is similar to the model developed in [5]; however, given the difference in the input data (multiple viewpoint 2D data in this case, single viewpoint 3D data in [5]), the processing needed to achieve such similar representation is different. In particular, since the skeletal tracker exploited in our previous system needs RGB-D data, it is not possible to exploit the same algorithm using RGB data only. Furthermore, dealing with multiple viewpoints, an additional challenge needs to be faced: the association of the same person among the different views analyzed. This means that a novel processing pipeline should be developed, with the constraint of providing, for each person in the scene, a signature for re-identification which can be compared with those evaluated from a single RGB-D sensor: in other words, even though the multiple view RGB and the single view RGB-D systems are based on different algorithms, they need to provide the same type of output data, which is crucial in order to compare observations performed using different sensory systems.

Overall, the system proposed in this paper presents the following elements of novelty: (i) a method for merging together estimations of the body pose made from different viewpoints: multiple hypotheses for each viewpoint are considered, and the final merging phase is performed in the 3D space; (ii) the application of a matching technique that is capable of associating the correct body pose among different views when multiple people are found in the scene; (iii) the application of the re-identification method already developed for 3D data in this new context; (iv) extensive tests on a publicly available dataset for re-identification applications.

It is important to observe that the method for merging together multiple views is decoupled from the single-view pose estimation algorithm, in contrast to what is proposed in [6]. This second option is focused on boosting performance, since the final 3D location of each joint is evaluated directly from all the images. On the other hand, our approach offers the advantage of being modular (because any 2D single-view body pose estimation algorithm can be used), simpler and with a very low computational cost, as illustrated in Section 4.5. Even though our approach does not have access to the source image while determining the 3D locations of the image joints, it considers several possible body poses guessed by the skeletal tracker, each one "voting" for a specific 3D body configuration. The RGB body pose estimation algorithm employed in this work is the Part-Based Detector (PBD) proposed in [2].

Regarding re-identification, the availability of multiple video sources lets the system observe each skeleton keypoint from several angles. This aspect will be considered in the future developments of the system, because the availability of multiple views of the same keypoint can enrich the target description, e.g., the front and rear part of a person can be seen at the same time by two cameras facing each other. The same advantage is not available in the single RGB-D sensor system described in [5].

The paper is organized as follows: in Section 2, previous work related to skeletal tracking and re-identification is discussed; in Section 3, the multi-view approach to skeletal tracking is detailed, together with the re-identification system exploited for performing the experiments. Experimental results are reported in Section 4, while the concluding remarks are summarized in Section 5.

2. Related work

People re-identification in images is addressed by observing three main characteristics: color, texture and shape, either considered separately or mixed together. A comparison and evaluation of the most important works in the literature is reported in [7]. The methods which exploit global histograms in RGB or HSV space assume that people can be distinguished by looking at their main colors and they keep the same appearance from every point of view. One of the best color-based approaches in the literature divides the body of each target into smaller parts and evaluates multiple histograms, one for each part [7,8]. This method is simple and effective, but suffers from two main flaws: it fails when the same target is seen in different illumination conditions, and it is a global (or semi-global) method that is not able to describe the target in detail.

Texture-based and shape-based approaches usually make use of local features and exploit descriptors evaluated on a set of keypoints to generate the signature of a target. Performance is therefore strongly related to the characteristics of the set of descriptors selected, including the capability of the keypoint detector to select stable features. This approach is widely used in the literature [9–11] thanks to its superior capability of providing a detailed description of each target; moreover, it overcomes the two main drawbacks of the color-based approach previously discussed. Such approach was also used together with histograms [12]. However, approaches based on local appearance feature extraction are usually computationally heavy because many features have to be matched at every frame and many mismatches can occur.

Very recently, computer vision for robotics was revolutionized by the introduction of affordable high-resolution three-dimensional sensors, that generate color point clouds instead of images. Moreover, efficient skeleton tracking algorithms [13] have been released which provide 3D position and orientation of the body joints from 3D data; some works also exploit temporal information [14] by tracking the skeleton joints. This had a strong impact on a number of applications, including people re-identification — for example, approaches based on three-dimensional features were developed. This new type of features can include information about both color and shape, which can provide superior performance over standard 2D features; however, noise affecting the location of the point cloud elements is usually not negligible for low-cost sensors like the Microsoft Kinect: in this case, shape descriptors can provide performance worse than expected, thus global approaches are usually preferred [15–19]. In [15] authors propose to build a body model of each person by projecting texture from different views to a mummy-like 3D model, which is then used for merging together short-term tracks that are recognized to belong to the same person. The matching stage aims at finding correspondences between pairs of models using shape and color information in order to calculate their similarity.

To overcome the problem of keypoint detection and matching, we recently proposed an approach for fast and accurate re-identification based on RGB-D data [5] that exploits the body joints, extracted by means of a skeletal tracker, by taking them as keypoints on which features (e.g. SIFT) are evaluated. The high

Please cite this article in press as: S. Ghidoni, M. Munaro, A multi-viewpoint feature-based re-identification system driven by skeleton keypoints, Robotics and Autonomous Systems (2016), http://dx.doi.org/10.1016/j.robot.2016.10.006