A Multi-view Framework for Human Parsing in Human-Robot Collaboration scenarios

Matteo Terreran, Leonardo Barcellona, Daniele Evangelista, Emanuele Menegatti and Stefano Ghidoni
Department of Information Engineering, University of Padua, Italy
Email: {terreran, barcellona, evangelista, emg, ghidoni}@dei.unipd.it

***Abstract*—Perception plays a major role in human-robot collaboration tasks enabling the robot to understand the surrounding environment, especially the position of humans inside its working area. This represents a key element to ensure a safe collaboration, and several human representations have been proposed in the literature (e.g., 3D bounding boxes, skeletal models). In this work, we propose a novel 3D human representation derived from body parts segmentation, which combines high-level semantic information (i.e., human body parts) and volume information. Body parts segmentation is known as human-parsing in the literature, which mainly focuses on RGB images. To compute our 3D human representation we propose a multi-view system based on a camera network, where single-view body parts segmentation masks are projected into 3D coordinates and fused together, obtaining a 3D representation robust to occlusions. A further step of 3D data filtering also improves robustness to outliers. The proposed multi-view human parsing approach has been evaluated in a real environment in terms of global and class accuracy on a custom dataset, acquired to thoroughly test the system under various conditions. The experimental results demonstrate that the proposed system achieves high performance also in multi-person scenarios where occlusions are largely diffused.**

***Index Terms*—human parsing, RGB-D perception, human-robot interaction**

I. Introduction

Human-robot collaboration (HRC) is one of the most studied topics in the robotics community. High importance is particularly given to safety, especially in industrial environments where robots represent potential sources of danger for human workers: when humans and robots operate simultaneously, they may be working very close together and accidental collisions between them must be avoided.

A common solution to guarantee safety in human-robot collaborative tasks is based on vision systems such as camera networks and people tracking algorithms [1]. Such systems may exploit different representations to describe the human pose and motion within the scene. In [2], people recognized by the detection algorithm are represented by means of a single point such as the person's centroid; this solution is fast, but it does not provide enough information to the robot for avoiding possible collisions. A simple improvement can be the construction of a 3D bounding box around the person's centroid [3], which allows to describe also the human volume. Other common human representations are based on skeletal models, namely a set of joints connected by a set of links [4]. These models provide a detailed representation of the person, with the position of each joint known at each instant, but do not provide information on the actual volume of the person.

In this work, we address the problem of human estimation by proposing a novel 3D representation based on body parts segmentation. We aim to segment people into different fine-grained semantic parts (e.g. head, torso, arms and legs), a problem known as human parsing in the literature. Our proposed representation contains the semantic information of the body parts, which allows to know at any time the position of the person and his/her body parts. Moreover, it allows a good and more refined estimation of the person's volume compared to its 3D bounding box.

II. Multi-view Human Parsing

A detailed picture of the proposed approach is given in Figure 1. Our system relies on a network of RGB-D cameras to be robust to occlusions [5]. In the first stage, a 2D human estimation is computed on RGB-D frames from multiple points of view, acquired by means of the camera network. For each viewpoint, people in the scene are segmented with respect to their body parts by means of our human-parsing module, based on the SCHP [6] architecture; moreover, an object detector localizes people with a bounding box, used later on for refinement. Single-view body parts segmentation masks are projected from 2D to 3D to obtain segmented point clouds from each camera, which are then aggregated together to overcome possible occlusions. A multi-view refinement is also used at this stage to remove noise and outliers, exploiting the single-view bounding boxes computed in the first stage. The final output of our system is a 3D semantic representation of each person in the scene; two representations are available, a segmented point cloud describing the person's volume and a high-level representation made only of the centroids of the body parts, similar to the skeletal models commonly used.

The proposed multi-view human parsing system has been built upon a previous work that addresses people and skeletal tracking in multi-view camera systems, known as OpenPTrack [2]. For the human-parsing module we used the original SCHP implementation[1] in PyTorch, using the

The research leading to these results has received funding from the European Unions Horizon 2020 research and innovation program under grant agreement No. 101006732. Part of this work was supported by MIUR (Italian Minister for Education) under the initiative "Departments of Excellence" (Law 232/2016).

[1] https://github.com/PeikeLi/Self-Correction-Human-Parsing

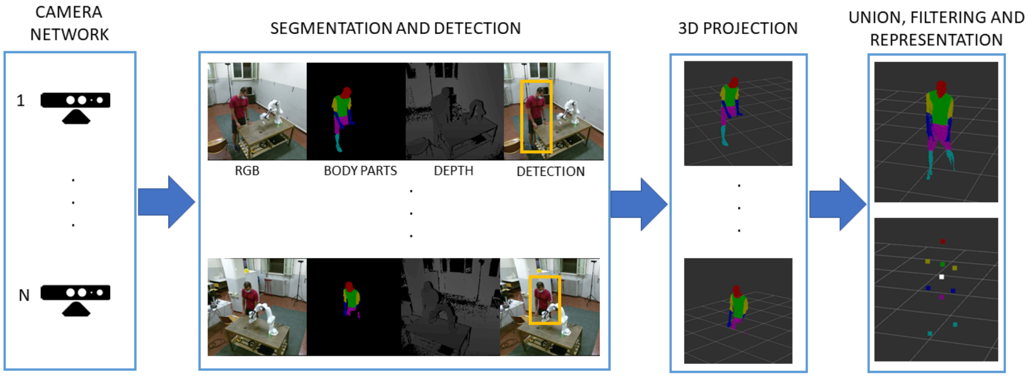
CAMERA NETWORK
1
N
SEGMENTATION AND DETECTION
RGB
BODY PARTS
DEPTH
DETECTION
3D PROJECTION
UNION, FILTERING AND REPRESENTATION

Fig. 1. An overview of our multi-view human parsing system. Body parts segmentation masks are computed from the RGB-D frames of each camera in the network, projected from 2D to 3D and then aggregated together; the final output of the system is a 3D semantic representation of the person in the scene.

TABLE I
CLASS AND GLOBAL PERFORMANCES ON THE CUSTOM DATASET, SUBDIVIDED PER TYPE OF SCENE. FIRST COLUMNS SHOW IOU PER CLASS. LAST COLUMNS SHOW THE GLOBAL PERFORMANCE.

Type of scene	Head	Torso	Upper arms	Lower arms	Upper legs	Lower legs	Background	GA	AP	F1	mIoU
Simple	68.5	77.2	56.4	36.4	71.6	63.2	99.8	99.1	83.2	79.6	67.5
Occlusions	75.2	75.1	50.4	42.4	68.7	50.7	99.3	99.1	76.7	75.5	62.8
Crowd	66.2	63.7	50.3	42.4	68.7	50.7	98.1	97.7	75.4	76.2	62.9
Crowd + occ.	74.8	72.4	53.7	37.7	66.6	51.2	98.8	98.4	78.7	77.4	65.0
Average	71.3	72.3	53.0	37.1	68.2	53.0	99.1	98.8	78.6	77.4	64.8

pretrained network weights provided by the authors after training on the Pascal-Person-Part dataset [7].

The multi-view human parsing system described so far has been evaluated on a custom dataset that was created on purpose. The dataset is composed of RGB-D frames acquired from multiple points of view using a camera network of Microsoft Kinect One sensors. The dataset includes 129 RGB-D frames from 3 different points of view, acquired under different conditions and levels of difficulty (i.e., one or more people, presence of strong occlusions, presence of a moving robot manipulator) to test and analyse the proposed approach in various scenarios. All the acquired frames have been manually annotated using the *Django Labeller*[2] image labelling tool.

A detailed analysis of the performance of our multi-view approach on each semantic class is given in Table I. Our method shows good performance on the classes *Head* and *Torso* in terms of mIoU, achieving good results even in the case of occlusions. The most critical class is *Lower arms*, for which we achieve low performance even in fairly simple scenarios. However, this result depends very much on the performance of human-parsing models, which in general struggle on such class.

III. CONCLUSIONS

In this work, we proposed a multi-view human parsing system capable of estimating a semantic 3D volume of people in a scene. Considering human-robot collaboration scenarios, our representation presents several advantages with respect to the representations commonly adopted such as bounding boxes and skeletons: flexibility, robustness as well as combining semantic and volume information, useful for implementing human collision avoidance strategies. Experiments on our custom dataset demonstrated how our multi-view approach helps to achieve high segmentation accuracy on scenes of various difficulty levels, such as in the case of strong occlusions or many people in the scene.

REFERENCES

[1] R.-J. Halme, M. Lanz, J. Kämäräinen, R. Pieters, J. Latokartano, and A. Hietanen, “Review of vision-based safety systems for human-robot collaboration,” *Procedia CIRP*, vol. 72, pp. 111–116, 2018.

[2] M. Munaro, F. Basso, and E. Menegatti, “Openptrack: Open source multi-camera calibration and people tracking for rgb-d camera networks,” *Robotics and Autonomous Systems*, vol. 75, pp. 525–538, 2016.

[3] M. Terreran, E. Lamon, S. Michieletto, and E. Pagello, “Low-cost scalable people tracking system for human-robot collaboration in industrial environment,” *Procedia Manufacturing*, vol. 51, pp. 116–124, 2020.

[4] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “Openpose: realtime multi-person 2d pose estimation using part affinity fields,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[5] A. Saviolo, M. Bonotto, D. Evangelista, M. Imperoli, E. Menegatti, and A. Pretto, “Learning to segment human body parts with synthetically trained deep convolutional networks,” *CoRR*, 2021.

[6] P. Li, Y. Xu, Y. Wei, and Y. Yang, “Self-correction for human parsing,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[7] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, “Detect what you can: Detecting and representing objects using holistic models and body parts,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1971–1978.

[2] https://github.com/Britefury/django-labeller