# Multi-view Human Parsing for Human-Robot Collaboration

# Multi-view Human Parsing for Human-Robot Collaboration

Matteo Terreran, Leonardo Barcellona, Daniele Evangelista and Stefano Ghidoni

***Abstract*— In human-robot collaboration, perception plays a major role in enabling the robot to understand the surrounding environment and the position of humans inside the working area, which represents a key element for an effective and safe collaboration. Human pose estimators based on skeletal models are among the most popular approaches to monitor the position of humans around the robot, but they do not take into account information such as the body volume, needed by the robot for effective collision avoidance. In this paper, 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. To compute such body parts segmentation, also known as human parsing in the literature, we propose a multi-view system based on a camera network. People body parts are segmented in the frames acquired by each camera, projected into 3D world coordinates, and then aggregated to build a 3D representation of the human that is robust to occlusions. A further step of 3D data filtering has been implemented to improve robustness to outliers and segmentation accuracy. The proposed multi-view human parsing approach was tested in a real environment and its performance measured in terms of global and class accuracy on a dedicated dataset, acquired to thoroughly test the system under various conditions. The experimental results demonstrated the performance improvements that can be achieved thanks to the proposed multi-view approach.**

## I. INTRODUCTION

Human-robot collaboration (HRC) aims to a close and direct collaboration between robots and humans to reach higher productivity and ergonomics thanks to the synergy between human intelligence and robot mechanical power [1]–[3]. Especially in industrial environments, robots are regarded as potential sources of danger: standard robotic cells have a fixed barrier to prevent the human from getting in contact with working machines. A first step toward human-robot collaboration is to remove any physical system separating the working environment of humans and robots, decreasing the amount of space and costs for safety barriers, but this requires alternative methods to ensure the safety of human workers. When humans and robots operate simultaneously, they may be working very close together as in Figure 1, 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 [4]. Such systems may exploit different representations to describe the human pose and motion within the scene. In [5], people recognized by the detection algorithm are represented by means of a single point such as the person's centroid; this allows a fast detection and message passing to the robot, but it is also less informative for the robot if it has to avoid possible collisions. A simple improvement can be the construction of a 3D bounding box around the centroid of the person [6], which allows to describe also the human volume: with this information the robot can then more easily avoid the space occupied by the person. However, how to choose the size of such a bounding box is not trivial, as it must be able to describe a large set of people (tall or short, thin or thick) without overestimating or underestimating the actual volume of the person. Other common human representations are based on skeletal models, namely a set of joints connected by a set of links [7], [8]. Representing the person by means of his/her skeleton has the advantage of giving a more detailed and useful description for human-robot collaboration, since it describes not only the position of the person but also the position and orientation of all the limbs (e.g. arms, hands). However, this is a very schematic representation: joints and links do not provide information about the volume of the person, thus limiting the possibility for the robot to avoid collisions in very close collaborative tasks.

Fig. 1. Example of human-robot collaboration scenario acquired with a multi-view camera setup. Overlaid on each image, the segmentation output of our multi-view human parsing system.

In this work, we address the problem of human estimation by proposing a novel representation based on body parts segmentation as shown in Figure 1. Our proposed representation combines the advantages of the human representations described above. On the one hand it is a representation that contains the semantic information of the body parts like skeletal representations with human joints, which allows to know at any time the position of the person and his/her body parts. On the other hand, it allows a good and more refined estimation of the person's volume compared to bounding box representations; this can be particularly useful in human-

All the authors are with the Intelligent Autonomous Systems Laboratory (IAS-Lab), Department of Information Engineering (DEI), University of Padova, Via Ognissanti 72, 35129, Padova, Italy. {terreran, barcellona, evangelista, ghidoni}@dei.unipd.it

robot collaboration scenarios as an input for motion planning algorithms to perform accurate collision avoidance. In particular, we aim to segment people into different fine-grained semantic parts such as head, torso, arms and legs. In the literature, such a problem is known as human parsing and allows to obtain a detailed representation of the person useful for high-level tasks such as recognition of human actions and gestures. The main state-of-the-art algorithms for human parsing tasks are based on deep learning networks [9]–[11], and focus on human body parts segmentation of people in RGB images. To the best of the authors' knowledge, no deep learning architecture has been proposed to deal with human parsing directly on 3D data (e.g, point cloud); this led us to develop a multi-view system to build a 3D semantic representation by exploiting the human parsing results from multiple points of view.

Our approach relies on a network of RGB-D cameras to be robust to occlusions [12]. The images acquired by each camera are segmented to recognise people and their body parts, and then projected from 2D to 3D exploiting the depth information acquired. The segmented 3D body parts from each viewpoint are then aggregated together to obtain a semantic 3D representation (i.e., segmented pointcloud) of the people in the scene, where the semantic information corresponds to the body parts. A 3D filtering method based on the multiview information is also proposed, which further improves the accuracy of the semantic 3D representation by removing noise and outliers. Popular datasets for human parsing tasks are composed of RGB images [13], [14] or synthetic RGB-D images [15], but none of them contains the same scene viewed from different points of view. Therefore, to evaluate our approach, a multi-view RGB-D dataset with human parsing manual annotations has been created on purpose[1], including various situations of increasing difficulty (single person, occlusions, crowded scenes, moving robot).

Summarizing, the work presents 3 main contributions: (i) a unified approach for people detection based on human body parsing and 3D projection; (ii) a 3D segmentation refinement step which exploits multiple views to reject misclassified points; (iii) a new multi-view RGB-D dataset with human parsing annotations, that can be used for testing algorithms in a real scenario to further drive research in this field. The remainder of the paper is organized as follows. Section II reviews the works related to human parsing and people detection. In Section III the main elements of our system are described in details, while in Section IV the proposed system is thoroughly evaluated on our dataset. Finally, in Section V, conclusions are drawn and future directions of research identified.

## II. RELATED WORKS

Human detection and tracking are crucial elements for human-robot collaboration, since the robot must be capable of perceiving the surrounding environment in order to identify potentially dangerous situations for human workers. Several representations can be adopted to describe human position and motion, depending on the particular type of collaborative operation.

In simple forms of collaboration, humans and robots share the same workspace but not at the same time: when a human enters this area, the robot must be stopped. In such a scenario, it is important to know the position in the scene and the volume of the person, using for example volumetric representations [16] or 3D bounding boxes [6]. For more direct collaborations, safety is ensured by maintaining at least a protective separation distance between human and robot at any time. Skeletal representations allow to monitor the distance of the robot from the various joints of the person's skeleton [17], [18]. In [19] a volumetric voxel-grid representation derived from skeletons is used to prevent potential robot collisions with humans, while in [20] human occupancy is represented in terms of convex volumes computed from skeleton joint positions. In [21] authors propose a human body representation made of 3D primitive shapes (e.g. spheres, cylinders), combining human skeleton detection with body parts semantic segmentation: given an RGB-D image, a set of shapes and parameters is estimated using both 3D skeleton joints and segmentation masks of each body parts (e.g. legs, arms).

In this work, we also investigate the use of body parts segmentation to compute a 3D representation of the human body, but without using any additional information on the human pose such as skeletons. Moreover, while in [21] authors focus on a single camera setup, we propose the use of a multi-view system to compute the 3D human body volume exploiting contributions from different viewpoints; this allows to improve accuracy and robustness to occlusions, typical of scenarios such as human-robot collaboration tasks due to movements of the robot or objects handled.

According to the literature, such representation based on the segmentation of body parts is known as human parsing [13], that is the task of segmenting a human figure in an image into different fine-grained semantic parts such as head, arms and legs. Several ideas have been proposed in the literature about how to define such semantic classes. For example, JPPNet [13] segments human body parts based on clothes. Other conventions, such as the one used in the Pascal-Person-Part [14] dataset, consider several classes of interest that are independent of the clothes worn and close to the human model also considered in skeleton-based algorithms (i.e., *Head*, *Torso*, *Upper arms*, *Lower arms*, *Upper legs*, *Lower legs*); in our work, we focus on such convention due to its close relation with the skeletal models commonly considered in the human pose estimation for human-robot collaboration.

Nowadays, state-of-the-art performance on the Pascal-Person-Part are achieved by deep learning architectures such as SCHP [9] and CDCL [10]. The former builds upon the model proposed by [22], which merges the segmentation information with a deep edge extractor module, proposing a cyclically learning scheduler to improve the model performance by progressively refining the noisy labels during

[1]Available at http://robotics.dei.unipd.it/

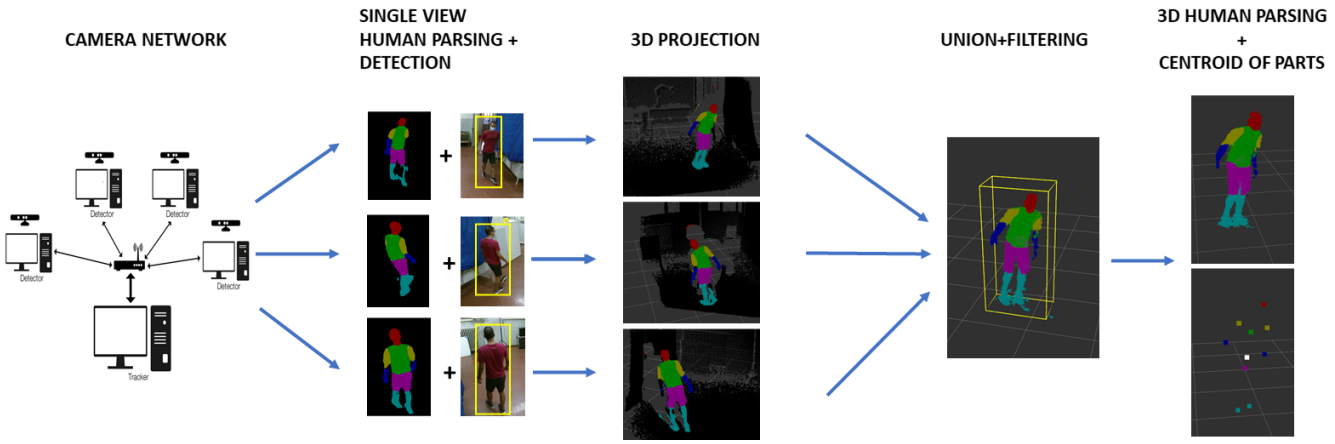


Fig. 2. An overview of our multi-view human parsing system. First, body parts segmentation masks and people bounding boxes are computed from the frames of each camera in the network; then single-view segmentation masks are projected from 2D to 3D and aggregate by means of our 3D filtering method. The final output of our system is a 3D semantic representation of the person, from which also a more high-level representation made of the centroids of the various body parts can be derived.

the training stage. The latter, in addition to the segmentation into body parts, also provides in output the skeleton of the person. The segmentation mask is obtained by training the model on a synthetic dataset, while the skeleton information allows to reduce the gap between synthetic and real data and to make the model more robust to real data. The idea of helping segmentation through the human pose has been also proposed in other works [11], [23], [24]. For example, the WSHP architecture [11] extracts the pose of a person, finds a set of stored segmentation masks with a similar pose, gets the average of them, and finally refines the mask generated with an encoder-decoder neural network that also takes as input the 3D pose of the person. A comparison between such architectures will be reported in Section IV. From such comparison, we selected the best candidate to be implemented in our system.

## III. METHODS

In this section, we provide a detailed description of the main parts of our system. A schematic representation of the proposed pipeline is shown in Figure 2, highlighting the main steps involved. In the first stage, a 2D human estimation is computed on RGB-D frames from multiple points of view, acquired by means of a camera network; for each viewpoint, people in the scene are segmented with respect to their body parts by means of our human-parsing module while 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 by means of the single-view bounding boxes. 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.

### A. Camera network system

We developed our multi-view human parsing system upon a previous work that addresses people and skeletal tracking in multi-view camera systems. In particular, we rely on OpenPTrack [5] and its user-friendly calibration procedure to setup and quickly calibrate the camera network. OpenPTrack[2] is a scalable and distributed multi-camera people tracking system with support for a heterogeneous set of cameras and 3D sensors (e.g., Asus Xtion, Microsoft Kinect One, Intel Realsense). It allows to perform people tracking within a network of RGB-D sensors by distributing people detection and centralizing the tracking process: each sensor is directly attached to a computer which analyzes the data stream and performs people detection; only the detections are sent through the network, in order to be merged at the tracking level after being referred to a common reference frame by means of calibration data, describing the pose of each camera within the network. Calibration data are obtained by means of a calibration procedure with checkerboards developed in ROS[3], the Robot Operating System.

Different people detection and tracking solutions are available in OpenPTrack, representing people either by means of a single centroid or with a skeleton. In our work, we propose a novel representation based on body parts segmentation which, unlike the skeleton, allows to describe also the volume of the person. Since it is based on segmentation, our representation is more robust to possible occlusions compared to OpenPTrack skeletons, as it will be shown in Section IV. Pose estimation methods based on skeletal models estimate the position of few points of interest (i.e., the skeleton joints), which can be quite inaccurate when the person is not entirely visible. Using semantic segmentation instead, we provide a pixel-level classification of the input highlighting each single body part and its pixels; only if a given body part is visible we can obtain the corresponding

[2]http://openptrack.org/
[3]https://www.ros.org

segmentation, thus reducing cases of incorrect predictions and improving the final 3D human representation obtained by combining information from multiple viewpoints. Our system relies on OpenPTrack essentially for network calibration, data exchange, and for collecting data from cameras through ROS; the segmentation-based approach developed is independent of OpenPTrack, although easy to be integrated into such framework in the future.

### B. Human parsing module

Our human parsing module is based on the SCHP [9] network, which proved to be the best candidate among the various models considered as reported in Section IV-A. The SCHP network, derived from CE2P architecture [22], is composed of a feature extraction backbone based on ResNet-101 and three main modules: context embedding module, high-resolution embedding module and edge perceiving module. The first module aims to extract global context information from the input image and involves a pyramid pooling module to generate features at multiple scales. Such features are then concatenated with low-level features from the backbone in the high-resolution module, which aims to recover lost details and provide output with high-level semantic and high-resolution spatial information. Finally, the edge perceiving module aims at learning the person contours to further refine the final prediction; multi-scale edge maps are computed by means of learned convolutions from the low-level feature maps of the backbone network, and then concatenated together into the output of the high-resolution module to predict the pixel-level human parts. Building upon this architecture, SCHP's authors propose a cyclical training procedure to progressively refine the noisy labels in the data, which further improves generalization performance and robustness of the model on several datasets. In our work, we used the original SCHP implementation[4] in PyTorch, using the pretrained network weights provided by the authors after training on the Pascal-Person-Part dataset.

### C. Multi-view refinement

In the last stage of our system, the detections coming from all the views are fused together. This is done by projecting the segmentation masks provided by each view from 2D to 3D and concatenating them together to obtain a 3D representation (i.e., point cloud) of each person in the scene including body parts semantic information. A common problem which arises in segmentation tasks is related to the false positives and false negatives: in both cases the segmentation model predicts a wrong class with respect to the ground truth. This leads to segmentation masks that may not perfectly match the edges of the subject, which is further amplified when projecting in the 3D space the segmentation output because mislabelled pixels could be projected onto surrounding objects, creating 3D clusters of human body parts that can confuse the robot.

Moreover, in a human parsing task the objective is to segment all the body parts in an image without distinction between the various instances of whole persons. This is a limitation in the case of human-parsing applications for human-robot collaboration scenarios, where the presence of robots and objects makes it difficult to recognize the number of people in the scene: due to occlusions, a single person can be segmented into a few distant body regions, making it difficult to know which person instance they belong to.

To solve both of the problems above, we propose a 3D refinement approach based on people detection and box filtering. We exploit the detection model YOLOv4 [25] to predict a 2D bounding box for each person in the scene; a 3D bounding box is then computed for each detected person considering both its 2D bounding box and the human parsing segmentation output. More in detail, we define $\mathcal{P}$ as the set of points $p$ such that:

$$p = [x, y, z, label, id] \in \mathbb{R}^5 \,,$$

where $x$, $y$, $z$ are 3D space coordinates, $label$ represents the segmented body part and $id$ is the instance number of the detected person.

We also define $\mathcal{B}$ as the set of boxes $b$ such that:

$$b = [p_{min}, p_{max}, p_{mean}] \,,$$

where $p_{min}$, $p_{max}$, $p_{mean} \in \mathcal{P}$ and all correspond to the same person instance (i.e., they all have the same *id* value). The objective is to find a set of boxes $\{b_1, \ldots, b_N\}$ where $N$ is the number of detected people in the scene and $b_i$ is a 3D bounding box enclosing all the segmented points of the i-th person, as depicted in Figure 3.

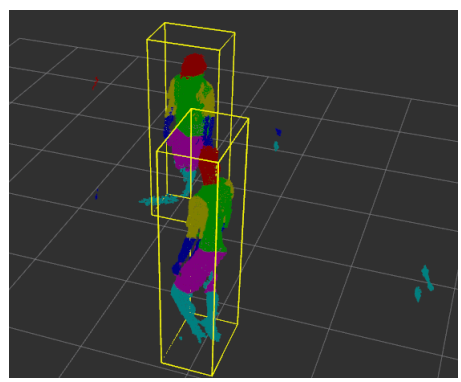

Fig. 3. Example of the 3D bounding boxes computed for the people detected in the scene, enclosing all the segmented body parts of each person. The colored points outside the boxes are noise that will be filtered out.

The procedure is shown in Algorithm 1. The inputs are the human parsing segmentation masks, the depth image, and the detections extracted from the people detector. Given the human parsing segmentation mask, we extract the contour of each body parts and compute the corresponding center of mass. If the center of mass is outside the contour (e.g., in the case of crossed arms), the function *iterateToCenter()* searches iteratively for a point belonging to the contour along an axis passing by the center of mass; the point found is then considered as the new center of mass in the following calculations. The $z$ coordinate of a point is computed as

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

**Algorithm 1:** Single view box creator

```
input  : hpImage, depthImage, detections
output: boxes
boxes ← null;
points ← null
contours ← findContours(hpImage)
foreach contour in contours do
    point ← contour.centerOfMass()
    if not(isInsideContour(contour, point))
     then
        point ← iterateToCenter(contour,
          point)
    end
    point.z ← meanDepth(depth, point.x, point.y)
    point.id ← id
    id ←
     findNearestDetectionIndex(detections,
     point)
    if id != -1 then
        point.label ← hpImage[ point.y, point.x ]
        point ← transformToWorld(point)
        points.push_back(point)
    end
end
for i=0 to i < max(ids) do
    mean ← findMean(points, i)
    points ← removeOutliers (points, mean, i)
    min,max ← findMinMax (point, i)
    mean ← (min + max)/2
    box ← [ min, max, mean ]
    boxes.push_back(box)
end
```

the mean of the neighbours' points in the depth image. The function *findNearestDetectionIndex()* returns the index of the detection containing the center or, if no detections are found, it returns the value $-1$. In the case of overlapping detection boxes, the index of the one with the nearest center to the box is returned. Finally for each set of points, identified by their corresponding $id$, we compute $p_{min}$, $p_{max}$ and $p_{mean}$ coordinates. The points too far from the mean are removed. This constraint is relaxed if the points are labelled *Lower legs* or *Lower arms*, because they tend to be more distant in common human poses. The boxes created from each view are then merged associating the mean positions to the nearest one if the distance is less than a given threshold. The result is a set of 3D bounding boxes, one for each person, built around the semantic 3D representation of the people detected in the scene as depicted in Figure 3. Points labelled as body parts outside these 3D bounding boxes are actually outliers, which can then be easily identified and removed. Thanks to this multi-view refinement, which exploits segmentation masks and 2D bounding box from multiple viewpoints, we can further improve the accuracy of our semantic 3D representation as demonstrated by our experimental results.

## IV. EXPERIMENTAL RESULTS

The multi-view human parsing system presented 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. For each sensor, intrinsics and extrinsics parameters were estimated using the calibration procedures implemented in OpenPTrack [5]. The dataset includes 168 RGB-D frames from 3 different points of view, acquired under different conditions (i.e., one or more people, presence of strong occlusions, presence of a robot manipulator) to test and analyse the proposed approach in various scenarios. All the acquired frames have been manually annotated using *Django Labeller*[5], a light-weight and open-source image labelling tool with support for many annotation shapes (e.g., polygons, boxes, oriented ellipses) and several utilities such as the automatic generation of polygonal outlines of objects identified by the user with a few clicks.

### A. Human parsing model selection

The selection of the body parts segmentation model has been done by testing three of the most accurate architectures for which pretrained weights were already available, namely: SCHP [9], CDCL [10], WSHP [11]. Such architectures have been tested on the Pascal-Person-Part dataset, the results are reported in Table I below. The performance of the three methods has been evaluated both in terms of Intersection Over Union (IoU) and the time required for inference (seconds). In particular, inference time was not mentioned for all networks in their corresponding papers; testing each model on a same dataset with the same hardware was important in order to choose the best candidate for our system among these models. For the tests we used a Nvidia Geforce GTX1650 GPU with 4 GB of graphic memory. The results show that CDCL is the best model in terms of accuracy, but the time needed for predicting the mask on a single image frame is much higher than the other architectures. Since this work addresses human robot collaboration scenarios, we finally chose SCHP; this network gives the best trade-off between IoU and inference speed, that is not irrelevant in such applications that aim to almost real-time requirements.

### B. Multi-view Human parsing evaluation

One of the main modules of our method is our multi-view refinement algorithm, introduced to improve the quality of the final 3D segmentation by removing any outliers due to other objects in the scene. To measure the improvement brought by our refinement strategy, we compared its performance with that of a simple procedure that combines the single-view pointclouds without using any filtering.

Results are reported in Table II, where the two strategies are evaluated in terms of mean Intersection over Union (mIoU), Global Accuracy (GA), Average Precision (AP), and F1-score. To improve clarity in Table II the results of

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

TABLE I

EVALUATION OF THE HUMAN PARSING ARCHITECTURES ON THE PASCAL-PERSON-PART DATASET [14]. FIRST COLUMNS SHOW IOU PER CLASS, WHILE THE LAST COLUMN REPORTS THE AVERAGE INFERENCE TIME FOR EACH MODEL.

| Model | Head | Torso | Upper arms | Lower arms | Upper legs | Lower legs | Background | Avg | Runtime per img(s) |
|---|---|---|---|---|---|---|---|---|---|
| SCHP | 87.41 | 73.80 | 64.98 | 64.70 | 57.43 | 55.62 | 96.26 | 71.46 | **0.127** |
| CDCL | 86.39 | 74.70 | 68.32 | 65.98 | 59.86 | 58.70 | 95.79 | 72.82 | 0.845 |
| WSHP | 87.15 | 72.28 | 57.07 | 56.21 | 52.43 | 50.36 | 97.72 | 67.60 | 0.609 |

segmentation into individual body parts have been combined into a single *Body* class, framing the comparison as a binary segmentation between *Body* and *Background*; for both classes the corresponding mIoU value is also reported.

From the obtained results we can see that our refinement procedure introduces a significant improvement in person segmentation: in all scene types of our dataset there is a remarkable improvement in terms of mIoU on the *Body* class. The performance improvement is mainly due to the ability of our method to recognize and remove all the segmentation masks that are not body parts (e.g., robot or objects in the scene) and filter out any outliers, such as those points along the contours of the segmented body parts that are projected onto the background when projecting from 2D to 3D. Indeed, the best improvement is achieved for the data acquired in the presence of occlusions, namely *Occlusions* and *Crowd + occ* in Table II, because in these scenarios the single-view segmentation mask predicted by SCHP tends to incorporate objects in the scene and the 3D filtering becomes fundamental to remove outliers. In terms of GA we do not obtain significant improvements because the class *Background* has a higher number of points correctly labelled, which makes the errors of body parts negligible.

TABLE II

EVALUATION OF OUR MULTI-VIEW REFINEMENT (MVR) METHOD IN DIFFERENT SCENARIOS. CENTRAL COLUMNS SHOW IOU PER CLASS, GROUPING ALL BODY PARTS INTO A SINGLE CLASS "BODY".

| Type of scene | MVR | Body | Bkgd | GA | AP | F1 | mIoU |
|---|---|---|---|---|---|---|---|
| Simple | ✓ | 72.40 | 99.47 | 99.21 | 92.13 | 91.86 | 85.93 |
| Simple | | 67.16 | 99.15 | 98.9 | 88.31 | 90.01 | 83.16 |
| Occlusions | ✓ | 64.91 | 99.33 | 99.07 | 88.84 | 89.19 | 82.12 |
| Occlusions | | 58.38 | 98.94 | 98.68 | 84.21 | 86.71 | 78.66 |
| Crowd | ✓ | 69.06 | 98.79 | 98.13 | 90.32 | 90.55 | 83.92 |
| Crowd | | 65.05 | 98.23 | 97.60 | 87.42 | 89.02 | 81.64 |
| Crowd + occ | ✓ | 66.84 | 98.81 | 98.31 | 88.80 | 89.41 | 82.33 |
| Crowd + occ | | 61.04 | 98.24 | 97.78 | 85.26 | 87.57 | 79.64 |
| Average | ✓ | 68.13 | 99.11 | 98.68 | 90.03 | 90.30 | 83.62 |
| Average | | 63.26 | 98.65 | 98.25 | 86.50 | 88.49 | 80.96 |

A detailed analysis of the performance of our multi-view refinement on each semantic class is given in Table III. 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 is a direct consequence of the performance of 2D human parsing models, which generally struggle with this particular category, as shown in Table I. Our 3D filtering method can only refine the segmentation masks by exploiting the multi-view information, but cannot segment body parts that were previously missed in the single-view segmentation.

### C. Comparison with People and Skeletal tracking

As pointed out in Section I, our human representation based on 3D body parts segmentation combines the advantages of other human representations typically considered in the literature, namely volume and semantic information. Moreover, our representation is very flexible and allows to easily derive the other representations. For example, by considering the centroid of each 3D body parts, we can estimate a schematic representation of the person similar to its skeleton's joints as shown in Figure 2. We can also construct an ellipsoid around the centroids of each body part by means of PCA, obtaining a representation similar to [20] and suitable for implementing collision avoidance algorithms in human-robot collaboration scenarios.

We rely on such an analogy between body parts' centroid and skeletons to evaluate the tracking performance of our system. In particular, for each type of scene in our dataset, we compared the trajectories described by the OpenPTrack people tracking algorithm with the trajectories of some centroids extracted from our representation. In OpenPTrack, the people tracking algorithm returns a centroid for each detected person positioned at belly height, while in our representation we consider the centroids of the classes *Head* and *Torso*, and the centroid of the pointcloud representing a whole person. We chose these centroids because they are generally aligned along the z-axis (perpendicular to the floor plane) with the centroid provided by OpenPTrack, thus allowing us to measure the tracking performance in terms of mean error along the $x$- and $y$-directions.

The result of this comparison has been obtained by performing 5 different test runs, that correspond to 5 different task settings: people walking on a straight line without any camera occlusion, people walking on a straight line with occlusions, people walking on a random path and 2 people walking randomly in the workspace. In the last setting, the error has been computed by considering each person instance in the scene. Results are given in Table IV, showing that, for all runs, the average error does not exceed $0.07$ meters even when occlusions with other people occur. Among the various types of centroids considered, the one corresponding to the *Head* class achieves the best results on the different

TABLE III

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 IN TERMS OF GLOBAL ACCURACY, AVERAGE PRECISION, F1 SCORE AND MEAN IOU.

| Type of scene | Head | Torso | Upper arms | Lower arms | Upper legs | Lower legs | Background | GA | AP | F1 | mIoU |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Simple | 76.94 | 79.80 | 58.88 | 47.49 | 79.65 | 68.79 | 99.47 | 99.21 | 85.05 | 83.75 | 73.00 |
| Occlusions | 78.88 | 76.31 | 58.77 | 40.73 | 65.99 | 51.27 | 99.33 | 99.07 | 79.97 | 79.62 | 67.32 |
| Crowd | 81.68 | 76.43 | 53.45 | 52.76 | 73.36 | 61.07 | 98.79 | 98.13 | 81.80 | 82.33 | 71.08 |
| Crowd + occ. | 80.42 | 75.21 | 58.04 | 49.25 | 65.37 | 51.35 | 98.81 | 98.31 | 79.26 | 80.32 | 68.35 |
| Average | 80.13 | 76.66 | 56.49 | 49.00 | 71.11 | 58.78 | 99.11 | 98.68 | 81.47 | 81.71 | 70.18 |

runs. Since our centroids are not the same as the one used by OpenPTrack, a detailed comparison to establish the accuracy of our trajectories is not possible; however, the experiment shows how the centroids derived from our representation can prove to be an interesting alternative for tracking people, especially in human-robot collaboration scenarios: either to monitor a safe distance between the robot and the human, or whether a large part of the person is occluded by the robot and only certain parts of the body are visible.

We also compared our semantic 3D body representation with the skeletal tracking solution implemented in OpenPTrack, just from a qualitative point of view, since our centroids cannot be directly expressed as a function of the skeletal joints. Some examples of such comparisons are depicted in Figure 4, showing OpenTrack's skeleton and our semantic representation on the same input data. Our representation proves to be more stable and robust to occlusions, unlike skeletons that require most of the person's body to be visible. For example, in the left image of Figure 4 with two people walking very close to each other, one of the people's skeleton is not detected while our semantic representation is computed for both people with good accuracy. Also regarding stability, our approach gives better results: the joints of the skeleton are often mistakenly found on objects in the scene resulting in a representation with wrong proportions, joint positions and directions of the links, as in Figure 4 on the right. Instead our representation is always very close to the person and so are the centroids that can be extracted from them.

TABLE IV

COMPARISON BETWEEN OUR APPROACH AND OPENPTRACK [5] PEOPLE TRACKING ALGORITHM. THE ENTRIES SHOW THE MEAN DISTANCE BETWEEN THE BODY PARTS CENTROIDS AND THE OPENPTRACK'S CENTROID IN DIFFERENT SCENARIOS [RESULTS IN METERS].

| Type of scene | Head | Torso | Body |
|---|---|---|---|
| Straight line | $0.030 \pm 0.019$ | $0.039 \pm 0.023$ | $0.034 \pm 0.033$ |
| Straight line + occ | $0.032 \pm 0.019$ | $0.039 \pm 0.020$ | $0.037 \pm 0.020$ |
| Random walk | $0.064 \pm 0.032$ | $0.056 \pm 0.039$ | $0.061 \pm 0.036$ |
| Two people - 1 | $0.056 \pm 0.032$ | $0.069 \pm 0.048$ | $0.076 \pm 0.041$ |
| Two people - 2 | $0.051 \pm 0.029$ | $0.058 \pm 0.034$ | $0.062 \pm 0.038$ |

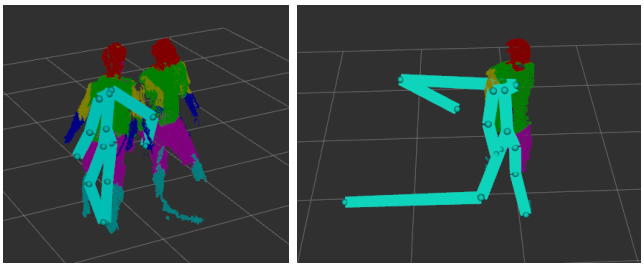

Fig. 4. Comparison between OpenPTrack skeletal tracking output (light blue) and our proposed 3D body parts segmentation. On the left, example of a person not detected by OpenPTrack but recognized with our semantic representation. On the right, example of an inaccurate skeleton estimated by OpenPTrack due to wrong joints positions.

## V. 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 and it also combines semantic and volume information, useful for implementing human collision avoidance strategies. The system is based on a camera network that provides RGB-D frames of the same scene from various viewpoints, and on a state-of-the-art human parsing network that segments the body parts of the people in each view; the single-view segmentations are then aggregated together by mean of a multi-view refinement method to obtain a semantic 3D representation of the people. In our experiments, we demonstrated how our multi-view refinement approach helps to achieve higher segmentation accuracy and provided a thorough performance analysis on the single body parts classes using a novel multi-view RGB-D dataset collected on purpose with scenes of various difficulty levels. Moreover, we compared our semantic representation with the representations used in people and skeletal tracking algorithms, highlighting the flexibility and robustness of our approach. As future research directions, we will investigate how to further improve the fusion of semantic information from individual views, and the possibility to combine our 3D representation with skeletal models to make them more robust.

## ACKNOWLEDGMENT

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).

## REFERENCES

[1] V. Villani, F. Pini, F. Leali, and C. Secchi, "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications," *Mechatronics*, vol. 55, pp. 248–266, 2018.

[2] E. Matheson, R. Minto, E. G. Zampieri, M. Faccio, and G. Rosati, "Human–robot collaboration in manufacturing applications: a review," *Robotics*, vol. 8, no. 4, p. 100, 2019.

[3] W. Kim, L. Peternel, M. Lorenzini, J. Babič, and A. Ajoudani, "A human-robot collaboration framework for improving ergonomics during dexterous operation of power tools," *Robotics and Computer-Integrated Manufacturing*, vol. 68, p. 102084, 2021.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] S. Ershadi-Nasab, E. Noury, S. Kasaei, and E. Sanaei, "Multiple human 3d pose estimation from multiview images," *Multimedia Tools and Applications*, vol. 77, no. 12, pp. 15 573–15 601, 2018.

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

[10] K. Lin, L. Wang, K. Luo, Y. Chen, Z. Liu, and M.-T. Sun, "Cross-domain complementary learning using pose for multi-person part segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 1066–1078, 2020.

[11] H.-S. Fang, G. Lu, X. Fang, J. Xie, Y.-W. Tai, and C. Lu, "Weakly and semi supervised human body part parsing via pose-guided knowledge transfer," *arXiv preprint arXiv:1805.04310*, 2018.

[12] 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.

[13] X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: Joint body parsing & pose estimation network and a new benchmark," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 4, pp. 871–885, 2018.

[14] 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.

[15] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 109–117.

[16] M. J. Rosenstrauch and J. Krüger, "Safe human robot collaboration—operation area segmentation for dynamic adjustable distance monitoring," in *2018 4th International Conference on Control, Automation and Robotics (ICCAR)*. IEEE, 2018, pp. 17–21.

[17] M. J. Rosenstrauch, T. J. Pannen, and J. Krüger, "Human robot collaboration-using kinect v2 for iso/ts 15066 speed and separation monitoring," *Procedia CIRP*, vol. 76, pp. 183–186, 2018.

[18] S. Yang, W. Xu, Z. Liu, Z. Zhou, and D. T. Pham, "Multi-source vision perception for human-robot collaboration in manufacturing," in *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*. IEEE, 2018, pp. 1–6.

[19] H. Liu and L. Wang, "Collision-free human-robot collaboration based on context awareness," *Robotics and Computer-Integrated Manufacturing*, vol. 67, p. 101997, 2021.

[20] M. Ragaglia, A. M. Zanchettin, and P. Rocco, "Trajectory generation algorithm for safe human-robot collaboration based on multiple depth sensor measurements," *Mechatronics*, vol. 55, pp. 267–281, 2018.

[21] R. Hachiuma and H. Saito, "Volumetric representation of semantically segmented human body parts using superquadrics," in *International Conference on Virtual Reality and Augmented Reality*. Springer, 2019, pp. 52–61.

[22] T. Ruan, T. Liu, Z. Huang, Y. Wei, S. Wei, and Y. Zhao, "Devil in the details: Towards accurate single and multiple human parsing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 4814–4821.

[23] X. Nie, J. Feng, and S. Yan, "Mutual learning to adapt for joint human parsing and pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 502–517.

[24] F. Xia, P. Wang, X. Chen, and A. L. Yuille, "Joint multi-person pose estimation and semantic part segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6769–6778.

[25] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.