Towards a holistic human perception system for close human-robot collaboration

Matteo Terreran[1], Leonardo Barcellona[1,2], Davide Allegro[1] and Stefano Ghidoni[1]

[1]*Intelligent and Autonomous System Laboratory (IAS-Lab), University of Padova, Italy*
[2]*Ph.D student in Artifical Intelligence, Politecnico di Torino, 10138 Torino, Italy*

Abstract
When considering close human-robot collaboration, perception plays a central role in order to guarantee a safe and intuitive interaction. In this work, we present an AI-based perception system composed of different modules to understand human activities at multiple levels, namely: human pose estimation, body parts segmentation and human action recognition. Pose estimation and body parts segmentation allow to estimate important information about the worker position within the workcell and the volume occupied, while human action and intention recognition provides information on what the human is doing and how he/she is performing a certain action. The proposed system is demonstrated in a mockup scenario targeting the collaborative assembly of a wooden leg table, highlighting the potential of action recognition and body parts segmentation to enable a safe and natural close human-robot collaboration.

Keywords
human-robot collaboration, human perception, body parts segmentation, action recognition

1. Introduction

Human-robot collaboration (HRC) aims at a close and direct collaboration between robots and humans to reach higher productivity thanks to the synergy between human intelligence on one side, and robot artificial intelligence and mechanical power on the other. This collaboration offers several benefits such as improved worker ergonomics, higher productivity, production flexibility and mass customization. Currently, many practical uses of human-robot collaboration in industry adopt a simplified form of collaboration, where humans and robots share the same workspace but at different times to guarantee human safety: if a person gets close to a robot that is in operation, the robot stops until the person moves away [1]. Such form of collaboration may introduce slowdowns in the production process and does not allow for various collaborative tasks in which person and robot must be in close contact (e.g., assembly or object-passing) or handle a large and heavy object together. However, this would often be the case: the best option in many assembly operations would require the human and the robot to work side-by-side to assemble an object composed of several components – typically the robot should assists the human by passing the tools or the parts, while the human completes the operations requiring dexterous manipulation. All such operations involve a close collaboration, that

9th Italian Workshop on Artificial Intelligence and Robotics (AIRO 2022)
matteo.terreran@unipd.it (M. Terreran); leonardo.barcellona@phd.unipd.it (L. Barcellona); davide.allegro.1@phd.unipd.it (D. Allegro); stefano.ghidoni@unipd.it (S. Ghidoni)
0000-0001-9862-8469 (M. Terreran); 0000-0003-4281-0610 (L. Barcellona); 0000-0003-3406-8719 (S. Ghidoni)

means coordination of actions and intentions between the robot and the human to maximize efficiency and to guarantee human safety. A crucial step to reach such adaptive behavior is the development of a perception system capable to monitor human position and activity within the workcell.

Many approaches have been proposed in the literature to estimate the position in the scene and the volume of the person, using for example volumetric representations [2] or 3D bounding boxes [1]. But when close human-collaboration collaboration is addressed, skeletal representations provided by human pose estimation algorithms are generally adopted, since they allow to monitor the distance of the robot from the various joints of the person's skeleton [3, 4]. Many human representations specialized for collision-avoidance can be further derived from skeletons: in [5] a volumetric voxel-grid representation derived from skeletons is used to prevent potential robot collisions with humans, while in [6] human occupancy is represented in terms of convex volumes computed from skeleton joint positions. However, such representations tend to overestimate a person's body size, and strongly depend on the output of pose estimation, which may be noisy or incomplete due to occlusions.

Recently, human pose has also been used as an input for human action recognition outperforming other approaches on popular action recognition datasets [7]. Action recognition is usually addressed focusing just on body information but, especially in collaborative assembly tasks, hands information can be very important to discriminate between very similar gestures (e.g., ok, stop) or actions where the body is mainly still (e.g., tightening a screw, assembling two interlocking pieces). However, obtaining accurate hand poses in these contexts is even more complicated than body pose estimation, leading many works to address hand pose estimation using ad-hoc setups with cameras that frame the hands very closely [8]; hands contains many joints to be estimated very close to each other, and are very easily occluded when objects or tools are manipulated.

In this work, we investigate AI perception methods to enable a closer collaboration in such applications where robot and human operator work in the same space at the same time on the same objects. To address this challenge, we propose an intelligent perception system capable of monitoring the whole robot workcell, providing information about the human workers: the system should be capable not only to detect human position and volume, but also to recognize what the human is doing (i.e., actions) and what he/she wants to achieve (i.e., intentions). Specifically, the perception system includes modules for pose estimation and action recognition, as well as a body parts segmentation module. Such module leads to several advantages compared to other works in the literature: (i) body parts segmentation provides an accurate estimate of the person's volume without depending on predefined geometric volumes or pose estimation results as other works in the literature; (ii) body parts segmentation allows to refine the output of the pose estimation module (e.g., by recovering missing joints), resulting in a refined representation of the human posture, especially with regard to hands.

The paper is organized as follows: in Section 2 each module of the system is presented in detail, while in Section 3 an experimental validation of the system is given. Finally, in Section 4 conclusions are derived and future developments of the system are illustrated.

2. Human perception system

The proposed system is based on a network of RGB-D cameras positioned around the robot workcell, providing information from multiple points of view to be robust to occlusions. All the cameras are calibrated both intrinsically and extrinsically, in order to express the information from each camera in a common coordinate reference frame (e.g., robot base). Each camera is attached to a processing device (e.g., PC) which analyzes the RGB-D data stream by means of AI-perception modules, providing mid-level information about people in the scene (e.g., pose estimation and body parts segmentation). The position of each camera with respect to the robot base frame is known. All the information is gathered by a main central PC which fuses them together to compute a unique 3D representation of the human worker describing his/her pose and volume; such representation is then used to compute high-level information about human activity (i.e., action and intentions). An overview of the system and its main AI-modules is shown in Figure 1, while in the following sections each module is described in detail.

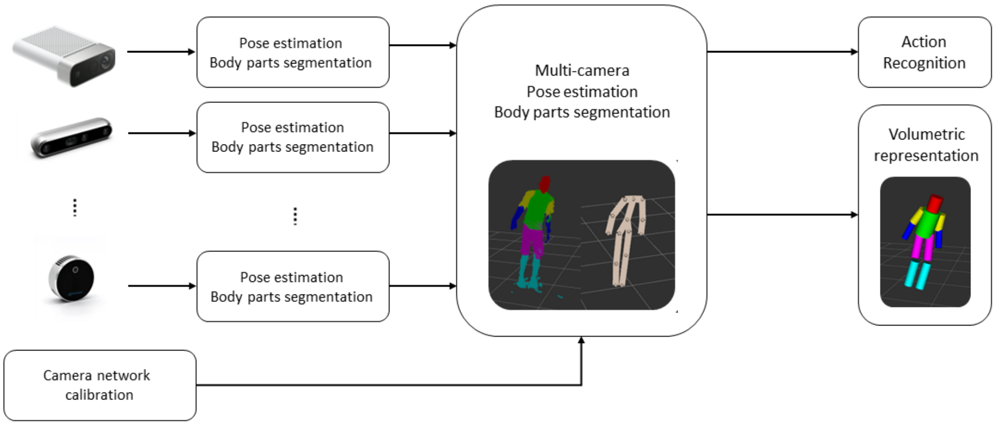

Figure 1: Overview of the human perception system. A dedicated processing node analyzes the RGB-D stream of each camera, computing human pose estimation and body part segmentation. All information is combined together by a central PC exploiting the camera network calibration, allowing to compute a volumetric human representation and to recognize human actions.

2.1. Camera network calibration

When dealing with multi-camera systems, it is very important to know precisely where each camera is with respect to the others. This information is the result of a calibration procedure, usually done by acquiring several images of a known pattern (e.g., checkerboard) from all cameras and by implementing an optimization process which allows to estimate unknown rigid transformations between sensors reference frame [9]. However, when considering a human-robot collaboration scenario we have the additional requirement to calibrate the camera network with respect to the robot base frame: in such a manner the robot can directly exploit the information from the perception system such as the position of the human worker, hence ensuring human safety by avoiding possible collisions.

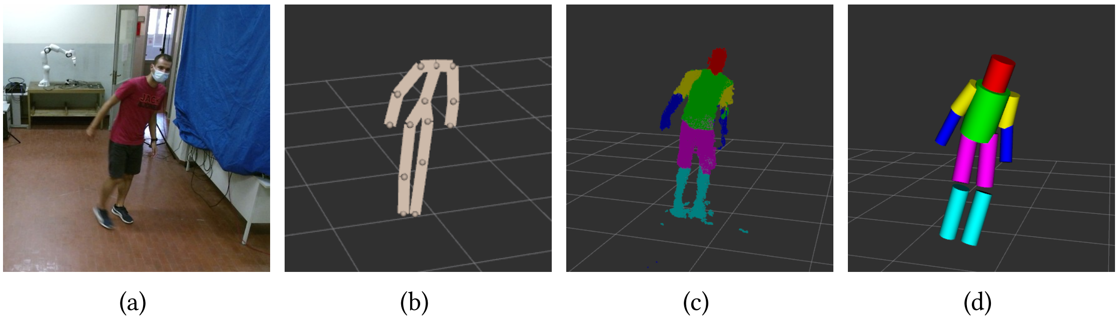

Figure 2: Different representation of a human body from the same scene: (a) RGB input image; (b) skeleton obtained as output of the pose estimation; (c) body parts segmentation, with a different color base on the pixel it represents; (d) a cylindrical representation obtained from (b) and (c).

To achieve such requirement, our perception system is calibrated by means of an hand-eye calibration procedure which allows to estimate the rigid transformation between the robot base and each camera in the network.

In particular, to calibrate the proposed system we rely on an iterative hand-eye calibration method based on non-linear optimization [10]. During the calibration phase, a planar checkerboard is mounted on the robot end-effector and moved around the workcell while acquiring images from all cameras; the pose of each camera with respect to the robot base frame is estimated as the rigid transformation minimizing the 2D euclidean distance between each reprojected 3D checkerboard corner on the image plane and the corresponding detected 2D corner. This method offers the main advantage of avoiding to directly use the transformation between the camera and the board and then it does not rely on Perspective-n-Point algorithms which could be unreliable with blurred images, that would negatively affect the whole calibration.

2.2. Human pose estimation and body parts segmentation

The human perception system is composed of several AI-based modules developed to provide a holistic understanding of the human worker, considering different types of information such as human pose and human volume. The human pose estimation module is based on the state-of-the-art OpenPose [11] detector which analyzes RGB input images and computes for each person in the scene a set of 2D points describing the joints of skeletal representation as in Figure 2b. OpenPose follows a bottom-up approach since it first extracts the joints position, without inferring the person they are related to, and then associates each joint to an identifier according to the person they belong exploiting part affinity field [12]. Such 2D points are then projected in the 3D space using the depth associated to the input RGB image and the camera intrinsic parameters, using a Kalman filter to merge the contributions from each camera.

Despite providing a very detailed information about the human pose and the position of the human worker within the workcell, the output of the pose estimation module does not include information like the volume occupied by the person which is also very important to implement human collision avoidance strategies in close collaborative tasks. For providing such complementary information, the proposed system includes also a human parsing module which runs in parallel to the pose estimation one. Such module aims to semantically segment an input RGB image assigning to each pixel a label representing a human body part (e.g., head, torso, arms, legs). The module is based on the SCHP [13] architecture, a state-of-the-art deep learning network for body parts segmentation on RGB images. The segmented output is then projected in the 3D world obtaining a labelled point cloud of the human worker (Figure 2c). Finally, the information from the pose estimation and human parsing modules are combined in a parametric human model representing each limb with a cylinder, as shown in Figure 2d. Each cylinder has direction and length obtained by the corresponding skeleton link, and radius computed from the labelled point cloud. Such cylindrical representation is useful not only for obstacle and collision avoidance, but also for direct interaction with the human. For example the information of the forearms can allow passing objects similarly to [14].

2.3. Human action recognition

The human action recognition module is based on an graph convolutional networks (GCN) [15] taking as input sequences of human body configurations (or human skeletons) computed by the previous modules. Sequences of skeletons provide a robust representation of the human movements free of any disturbances like external objects, lighting, and aesthetic differences of people (e,g., clothes or skin color). Therefore they represent an interesting source of information to achieve a robust and general action recognition, especially in collaborative scenarios where both human and robot are moving, and the human worker should interact with many objects and tools. In order to improve accuracy and robustness of the system, it relies on an ensemble of GCNs where each network is trained to recognize actions based on a different set of joints (e.g., body joints, hand joints, arm joints) and the final prediction is given by averaging all the networks' predictions [16]. In particular, the vision system recognizes the person's actions at various levels: a general classification of the type of action taking place (e.g., pick, place, request, hand to), and a finer recognition of the main direction of the movement and its intensity (e.g., small, medium, high intensity) useful for better characterize particular actions such as *pulling*, *pushing* or *pointing*.

3. Experimental validation

The proposed human perception system has been validated in a real scenario, that is the collaborative assembly of a small wooden table shown in Figure 3. In particular, the human operator and the robot work together to build a table composed of wooden and 3D-printed parts: the person performs the actions that require more manual dexterity, such as inserting parts (Figure 3a), while the robot assists the human by passing at the right time the parts that the human partner needs (Figure 3c).

The experimental setup is composed of a Franka Emika robot arm and a camera network of 4 RGB-D cameras (i.e., Microsoft Kinect V2) positioned at the four corners of the lab room, so as to observe the scene from multiple viewpoints and reduce the possibility of occlusions. Each camera is attached to a local PC equipped with a high-end NVIDIA GeForce RTX 2080 GPU, running both the pose estimation and body parts segmentation modules. All PCs are connected to a local network and send the outputs of the perception modules to a central PC through

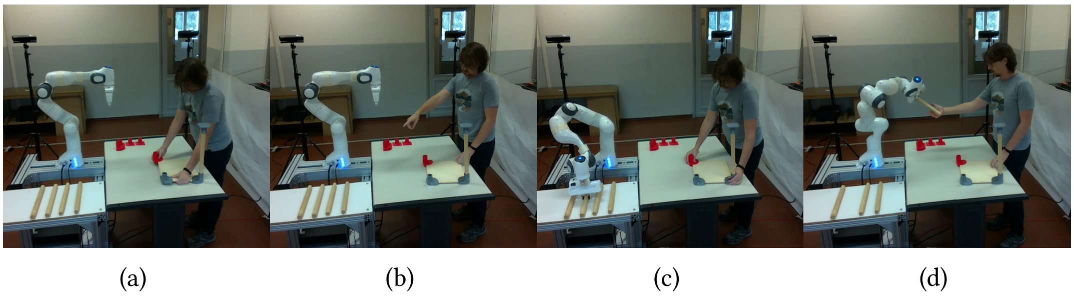

Figure 3: Example of a collaborative assembly task: (a) the operator performs the actions that require high manual dexterity, such as inserting parts; (b) the human worker requests a new object using a ”pointing” gesture; (c) the robot moves to pick the requested object while the human continue the assembly process; (d) the robot passes the object to the human partner, exploiting pose estimation and human parsing information to precisely localize the human hand and to avoid possible collisions.

ROS (Robot Operating System). Such central PC merges all contributions to obtain a unique representation of the person for each module (i.e., 3D pose and body parts segmentation), which are then used as input to the action recognition module.

The position of the human worker and his/her activities are constantly monitored thanks to the perception system developed, enabling an effective and intuitive human-robot interaction. When a new object is required during the assembly process, the worker can simply point to the object he/she wants to receive as in Figure 3b: the perception system recognizes the “pointing” action and the corresponding intention (i.e., the direction given by the arm), triggering the robot to move and pick the requested object. Once the robot has picked up the object, it moves in front of the operator at a safe distance to signal that it is ready to deliver the object. When the operator is ready to receive the object, he/she extends the arm with the hand open (i.e., “pass object” action) and the robot passes the object to the human partner, exploiting pose estimation and human parsing information provided by the perception system to precisely localize the human hand and to avoid possible collisions (Figure 3d).

4. Conclusions

In this work, the design of a AI-based perception system for close human-robot collaboration was presented. Special emphasis was placed on achieving effective and intuitive collaboration for the human operator through body parts segmentation and action recognition. The proposed system has been applied to a collaborative assembly task in a mock-up scenario, highlighting its potential for enabling a safe and natural human-robot collaboration in industrial scenarios.

References

[1] M. Terreran, E. Lamon, S. Michieletto, E. Pagello, Low-cost scalable people tracking system for human-robot collaboration in industrial environment, Procedia Manufacturing 51 (2020) 116–124.

[2] M. J. Rosenstrauch, 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.

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

[4] S. Yang, W. Xu, Z. Liu, Z. Zhou, 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.

[5] H. Liu, L. Wang, Collision-free human-robot collaboration based on context awareness, Robotics and Computer-Integrated Manufacturing 67 (2021) 101997.

[6] M. Ragaglia, A. M. Zanchettin, P. Rocco, Trajectory generation algorithm for safe human-robot collaboration based on multiple depth sensor measurements, Mechatronics 55 (2018) 267–281.

[7] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+ d: A large scale dataset for 3d human activity analysis, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1010–1019.

[8] T. Kobayashi, Y. Aoki, S. Shimizu, K. Kusano, S. Okumura, Fine-grained action recognition in assembly work scenes by drawing attention to the hands, in: 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), IEEE, 2019, pp. 440–446.

[9] M. Munaro, F. Basso, E. Menegatti, Openptrack: Open source multi-camera calibration and people tracking for rgb-d camera networks, Robotics and Autonomous Systems 75 (2016) 525–538.

[10] D. Evangelista, D. Allegro, M. Terreran, A. Pretto, S. Ghidoni, An unified iterative hand-eye calibration method for eye-on-base and eye-in-hand setups, in: 2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA), 2022, pp. 1–7. doi:10.1109/ETFA52439.2022.9921738.

[11] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, Openpose: Realtime multi-person 2d pose estimation using part affinity fields, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (2019) 172–186.

[12] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7291–7299.

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

[14] P. Rosenberger, A. Cosgun, R. Newbury, J. Kwan, V. Ortenzi, P. Corke, M. Grafinger, Object-independent human-to-robot handovers using real time robotic vision, IEEE Robotics and Automation Letters 6 (2020) 17–23.

[15] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, H. Lu, Skeleton-based action recognition with shift graph convolutional network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 183–192.

[16] M. Terreran, M. Lazzaretto, S. Ghidoni, Skeleton-based action and gesture recognition for human-robot collaboration, in: International Conference on Intelligent Autonomous Systems, Springer, 2022.