A Novel 3D Sensing Framework for Safety Monitoring in Human-Robot Collaboration Work Cells

Tarek Yahia[1], Kody Haubeil[1], Alexander Suer[1], Yongzhi Qu[2], Janet Dong[1], Xiaodong Jia[1]

[1] *Center for Intelligent Metrology & Sensing, Department of Mechanical & Materials Engineering, University of Cincinnati, Cincinnati, Ohio, 45219, USA*

yahiatk@mail.uc.edu, haubeiky@mail.uc.edu, suerad@mail.uc.edu, dongjg@ucmail.uc.edu, jiaxg@ucmail.uc.edu

[2] *Lab of Artificial Intelligence Powered Systems, Department of Mechanical Engineering, University of Utah, Salt Lake City, Utah, 84112, USA*

u6051628@utah.edu

ABSTRACT

This paper introduces a novel 3D sensing framework for real-time safety monitoring of Human-Robot Collaboration (HRC), aimed at injury risk reduction and enhancing worker safety. The framework uses data from a single RGB-D camera to generate a 3D human avatar. Human shape and pose are estimated using deep neural networks, which incorporate depth information and undergo 3D geometric transformations to determine accurate scale and translation. This process yields a reconstructed 3D mesh which captures the human's pose, shape, size, and location. Following 3D Human Pose Estimation (HPE), both the human avatar and the robot's digital twin are integrated into a shared virtual environment, enabling real-time monitoring of the HRC workspace. Results demonstrate effective reconstruction of 3D human geometry within HRC settings. By combining the reconstructed human surface mesh and real-time robot state in a single virtual environment, the system enables continuous, real-time monitoring of both the robot and the human agents.

1. INTRODUCTION

Robots have emerged as indispensable manufacturing tools in modern industrial environments due to their precision, speed, and ability to operate continuously with few obstructions to their productivity. As robotic technology has advanced, their presence in shared workspaces with human operators has become increasingly common. However, despite their technical capabilities, industrial robots remain entirely unaware of their surroundings. They are typically programmed to perform fixed tasks within a defined workspace, with little to no understanding of the dynamic presence of nearby humans. This lack of spatial awareness poses significant safety risks, making it necessary to confine robots to fenced-off cells or designated zones to prevent collisions with human operators. While effective for ensuring safety, such physical separation limits the flexibility and efficiency of HRC. As manufacturing moves towards more adaptive and collaborative workflows, there is a growing need for robots that can operate safely and intelligently alongside humans in shared environments.

For real-time monitoring of Human-Robot Interactions (HRI), machine vision represents the most viable and necessary technological direction. Vision-based systems, particularly those utilizing RGB-D cameras, offer a low-cost and easily deployable solution while retaining rich contextual information from the environment. Unlike other sensing modalities, visual data can be readily interpreted by safety engineers and data analysts, facilitating more informed assessments of safety risks. Traditional human monitoring methods, such as vision-based HPE, focus on estimating joint locations to construct 2D skeletal representations of humans (Fang et al., 2022; Sun et al., 2019). State-of-the-art object detection models, such as YOLOv11 (You Only Look Once, version 11) (Jocher et al., 2024), offer high-speed 2D human detection and are widely used for real-time perception tasks. However, these models provide only bounding boxes and coarse 2D localization. Some models extend this by 'lifting' 2D poses into 3D, leveraging mesh-based regressions to estimate 3D joint positions (Cao et al., 2019). While effective for basic activity recognition, these models are insufficient for applications requiring safety monitoring between robots and humans for three main reasons:

(1) Skeletal outputs lack volumetric detail, making them unreliable for estimating proximity in collision detection.

(2) 2D joints are predicted independently, resulting in a lack of kinematic regularization and, consequently, anatomically incorrect human poses.

(3) 3D joints are predicted in an arbitrary coordinate space, using a select joint (ex. mid-hip) as the origin. Without global translation, these poses lack real-world context, limiting their use in safety-critical applications.

Some approaches aim for more precise human detection in collaborative workspaces. (Mohammed et al., 2017) capture 3D point clouds with RGBD cameras and manually label human presence based on prior knowledge of the HRC workspace layout. (Kamezaki et al., 2024) track human movement using a system of three calibrated laser range finders, while (Patalas-Maliszewska et al., 2024) identify human and robot body parts from a top-down view using 2D CNNs. Despite their improvements, these methods are often environment-specific, require manual setup, and can produce unrealistic human poses, limiting their applicability for flexible and adaptive monitoring. To address these limitations, a full-surface mesh is required—offering a detailed and physically realistic 3D representation of the human body while incorporating depth information to enable accurate spatial localization.

Achieving human mesh reconstruction in real-time remains a significant challenge due to computational limitations in regularizing human kinematics. Several methods incorporate physics-based dynamic models for regularization, such as PhysMoP (Yufei Zhang et al., 2024) and PhysCap (Shimada et al., 2020) which enforce physics-based environment constraints for motion prediction and collision handling. Other methods, such as DynaIP (Yu Zhang et al., 2024) propose models which regularize human motion with acceleration data captured from wearable Inertial Measurement Unit (IMU) sensors. While these approaches improve physical plausibility, they rely on computationally intensive optimization methods. This high computational cost limits their feasibility for real-time applications in industrial HRC work cells.

The framework proposed in this paper is built entirely on a neural network architecture, enabling real-time performance without relying on computationally intensive physics-based models. This design enables fast and efficient inference suitable for industrial deployment. Unlike conventional 2D skeleton-based approaches, our method reconstructs the human as a full 3D surface mesh, capturing realistic body shape, size, and pose. Importantly, the reconstruction is expressed in real-world coordinates, enabling spatial reasoning with respect to other elements in the environment, including the robot. By integrating this representation with real-time data from the robot controller, the proposed framework creates a virtual model of the shared workspace. This allows continuous monitoring of HRI, including real-time tracking of human motion trajectory and distance from the robot, both of which will later be demonstrated in our results.

The rest of the paper is organized as follows. Section 2 introduces the HRC safety problem. Section 3 details the proposed framework. Section 4 presents the results and a corresponding discussion, while Section 5 summarizes the main conclusions.

2. Problem Statement and Related Works

According to International Federation of Robotics (IFR)-World Robotics 2023, it is reported that there are nearly 4 million industrial robots in operation worldwide, with approximately 10 percent of them being collaborative robots (cobots) (IFR, 2023). A National Institute of Occupational Safety and Health (NIOSH) report highlighted 61 robot-related fatalities between 1992 and 2015, with an expectation of further increase due to the increasing use of industrial robots and cobots in the US work environment (NIOSH, 2022). A recent study in (Lee et al., 2021) delved into 355 robot accidents documented by Korea Occupational Safety and Health Agency (KOSHA) between 2009 and 2019, revealing that 95% occurred in manufacturing businesses. Pinch and crush incidents accounted for 52% of the accidents, while impacts and collisions accounted for 36%, and the remaining 12% involved falls, flying objects, trips/slips, cuts, burns, etc. These findings align with US data reported in (Jiang & Gainer Jr, 1987). In terms of human injuries, the majority (31%) affected the hands and fingers, followed by 24% in the head and face, 22% in the neck and chest, 9% in the abdomen and back, 7% in the legs and feet, and 7% in the arms.

Power and Force Limit (PFL) and Speed and Separation Monitoring (SSM) are widely employed methods for mitigating collision injuries in HRI (Robla-Gómez et al., 2017). PFL enhances safety by regulating the force and power exerted by the robots (Aivaliotis et al., 2019), ensuring compliance with safety standards (ISO, 2013), and conducting risk assessments before HRI deployment. Recent advancements in PFL also cover the utilization of force sensors for collision detection (Magrini et al., 2015) and the development of lightweight robots to reduce impact (Hirzinger et al., 2000). On the other hand, SSM focuses on preemptive safety measures by monitoring relative speeds and maintaining appropriate distances between humans and robots before potential collisions occur (Byner et al., 2019; Campomaggiore et al., 2019; Marvel & Norcross, 2017). Key components in an SSM solution include 3D vision for position and speed monitoring, evaluation of human-robot separation (or minimum distance checks), and collision avoidance strategies alongside adaptive robot control.

To achieve SSM in HRC work cells, real-time estimation of the 3D human motion trajectory remains a challenge. While existing State of The Art (SoTA) deep learning methods can accurately predict 2D skeletons from traditional RGB images, accurate 3D motion trajectories still rely on expensive motion capture systems. This paper presents a cost-effective framework for SSM that estimates 3D human motion using a single RGB-D camera. It integrates 3D

vision technology with safety engineering principles and adaptive robot control to support collision detection and avoidance.

3. Proposed Methodology

3.1. SMPL

The SMPL (Skinned Multi-Person Linear) model is a parametric 3D human body model that represents realistic and pose-dependent body shapes using a low-dimensional set of parameters(Loper et al., 2023). It is widely used in computer vision and graphics for tasks like human pose estimation, motion capture, and avatar creation due to its differentiable formulation and compatibility with optimization and deep learning frameworks.

In SMPL, mapping a body from the rest pose $\overline{\mathbf{T}} \in R^{3\times6890}$ to a specific articulated pose involves a series of geometric transformations $M(\theta,\beta) \in R^{3\times6890}$ that are parameterized by the pose parameter $\theta \in R^{72}$ which represents the rotation of 24 body joints in axis-angle format, and the shape parameter $\beta \in R^{10}$. The geometric mapping process in SMPL involves two major steps.

Step 1: applying the blend shapes to the rest pose:

$$\mathbf{T}_p(\beta,\theta) = \overline{\mathbf{T}} + B_S(\beta) + B_P(\theta) \quad (1)$$

Where $B_S(\beta)$ and $B_P(\theta)$ are linear operations that compute the offset from the template pose $\overline{\mathbf{T}}$. It gives a new set of 6890 vertices that considers the shape and pose differences.

Step 2: rotating the $\mathbf{T}_p$ vertices to represent the current pose

$$M(\beta,\theta) = W(\mathbf{T}_p,\theta,\beta) \quad (2)$$

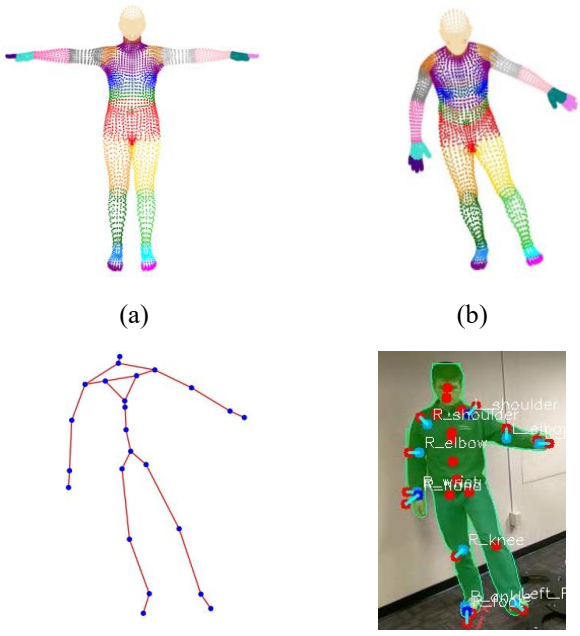
Where $W(\cdot)$ is a rotation matrix operation derived from Rodrigues' formula.

The mapped vertices from Eq.(**2**) represent the current pose of the human. Based on the mapped vertices, one can further obtain the 3D and 2D joint locations by using the following two equations.

$$\hat{X} = M(\theta,\beta)\cdot w,\ \hat{X} \in R^{3\times24} \quad (3)$$

$$\hat{x} = \pi(X) = s\cdot\Pi\left(Rot\cdot\hat{X}\right) + t \quad (4)$$

Where $\hat{X}$ is the 3D joints and $\hat{x}$ is the 2D joints corresponding to a given image. The parameter w in Eq. (**3**) is a pre-generated matrix. In Eq. (**4**), s and t denote the scaling and translation parameters needed to achieve the mapping. A graphical representation of these four mappings is illustrated in Figure 1.



(a) (b) (c) (d)

Figure 1. Intermediate geometric mappings of the SMPL model. (a) The template pose $\overline{\mathbf{T}}$; (b) The transformed pose $M(\beta,\theta)$ using Eq.(**1**) and (**2**); (c) The 3D joints given by Eq.(**3**); (d) the 2D joints given by Eq.(**4**).

3.2. Human Pose Estimation

The proposed method for 3D HPE is outlined in Table 1. A single-view RGB-D camera is deployed to monitor the HRC work cell, recording video at 30 frames per second (FPS). At each time step, an RGB image ***I*** and a corresponding depth map ***D*** are captured for further processing.

The data processing begins with estimating the SMPL model parameters using the pre-trained backbone network VIBE (Video Inference for human Body pose and shape Estimation) (Kocabas et al., 2020). It is a deep learning framework designed to estimate 3D human meshes from single-view RGB videos, leveraging temporal information across consecutive frames to produce more accurate and stable predictions compared to single-frame methods. VIBE uses a pre-trained pose detector to extract joint locations, which are then passed through a temporal convolutional network to regress the parameters of the SMPL model, including body shape β, body joint rotations θ and the scaling and translation factors s,t. In this study, we adopt the original backbone network without any fine-tuning.

Table 1. The proposed HPE algorithm.

At each time t, given an RGB image $\boldsymbol{I}$ and corresponding depth map $\boldsymbol{D}$, the following steps are followed to obtain 3D human pose.	
1	Estimate the SMPL model parameters β, θ, s, t using a pre-trained backbone model named VIBE (Kocabas et al., 2020). The input of the deep network is the RGB image I, the output is the SMPL model parameters.
2	Obtain the transformed vertices $M(\beta, \theta)$ *from* Eq.(1) and (2), the 3D joints $\hat{X}$ from Eq.(3), and the scaled 2D joints $\hat{x}$ from Eq.(4).
3	Augment the depth information to the scaled 2D joints $\hat{x}$ (see Figure 1-d) and create a measured 3D body joint $\hat{X}_m$ with proper scaling and location in the global coordinate system.
4	Estimate the scaling factor $s' \in R$ and translation vector $t' \in R^3$ by aligning the measured 3D joints $\hat{X}_m$ and the unscaled 3D joints from SMPL $\hat{X}$.
5	Apply the following transformation $s' \cdot M(\beta, \theta) + t'$ to the transformed vertices in step 2.
6	March to time $t + 1$ and repeat steps 1~5.

3.3. Robot Digital Twin Integration

The setup for integrating the collected RGB-D data, real-time robot controller data, and a computer to process and visualize the real-time monitoring is illustrated in Figure 3. An RGB-D sensor is used to capture human motion within a predefined work zone. The video data is transmitted to an industrial PC equipped with an NVIDIA GeForce RTX 2060 GPU. Using the proposed method described in Table 1, the video stream is processed into a 3D human avatar within a global coordinate system. Simultaneously, real-time robot joint angles are streamed from the robot controller to generate a digital twin, enabling synchronized robot motion in the virtual environment. As the entire pipeline is based on deep neural network computations, the system operates in near real-time with minimal latency.

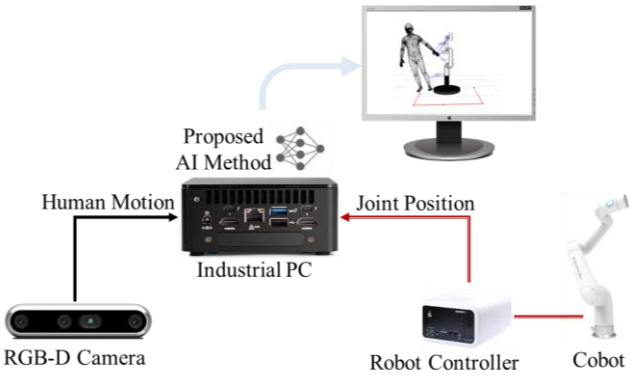


Figure 3. The hardware configuration for real-time HRC monitoring and robot digital integration.

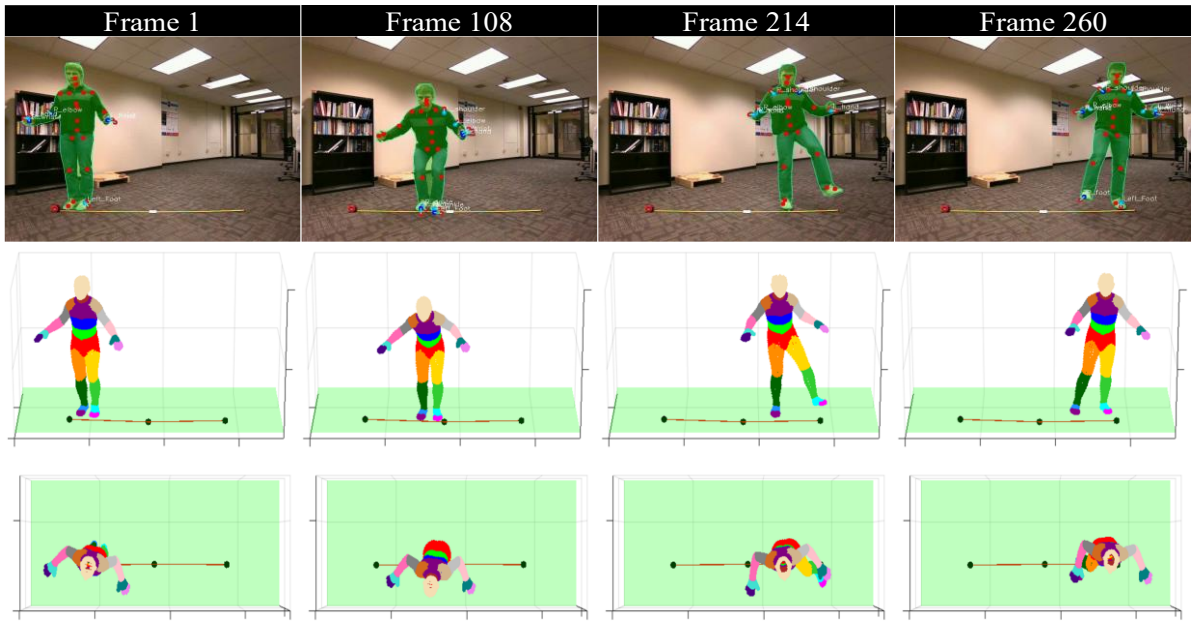


Figure 2. (Top row) Select frames from the RGB-D video collected in case study 1; (Middle row) Estimated front view using the proposed method; (Bottom row) Corresponding top view using the proposed method.

4. Results and Discussions

4.1. Case Study 1: Human Pose Estimation without Robot

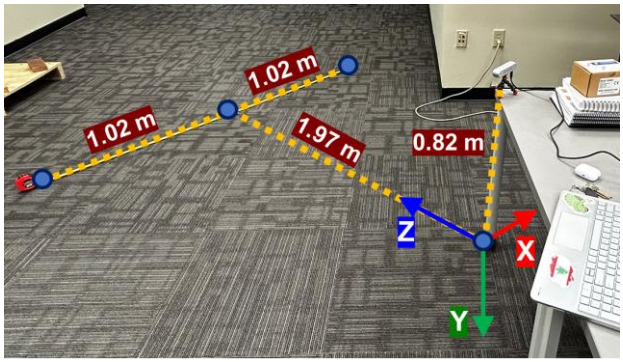


Figure 4. Experimental setup in case study 1.

The first case study involves a human walking along a straight path while performing various poses at different locations along the path (Figure 4 illustrates the setup designed to perform this experiment). RGB-D data was collected at 30 FPS using an Intel RealSense D455. Unlike traditional HPE tasks, which typically annotate human poses using 2D joints or bounding boxes, this study aims to localize a 3D avatar within a global coordinate system and estimate body joint positions relative to that global frame. Since global coordinates are required, the depth information collected from the RGB-D camera is key to recovering the 3D location of the human, and this task cannot be performed with RGB images alone. The framework processes only a small subset of the depth data—specifically, the pixels corresponding to the human's key points—allowing for real-time processing capable of matching the 30 FPS streaming rate. Median filtering is applied to this subset to remove outliers and improve the accuracy of the 3D localization. As the mesh vertices are regressed from a pre-generated human mesh template, they require no depth processing and undergo the same scale and translation transformations as the human key points.

The proposed method was utilized to process the RGB-D data and resulted in successful 3D mesh reconstruction of the human. Figure 3 shows the temporal progression of human motion, with distinct poses performed across the four displayed frames. The top row visualizes the 2D joints overlayed onto the image for each frame, while the middle and bottom rows demonstrate that our proposed method effectively reconstructed the 3D mesh of the human with accurate poses as seen in the corresponding RGB images from the first row. Notably, this method relies on input from a single RGB-D camera, which captures data of the person from one view, showing their front side. Despite this limitation, the top-down view of the 3D avatar reveals that the proposed method can realistically construct the entire shape and pose of the human figure.

To evaluate the accuracy of our proposed method, we compared the estimated human trajectory against the trajectory predicted by YOLOv11 (Jocher et al., 2024) as illustrated in Figure 5. Specifically, the YOLOv11n-pose model is used to predict the 2D human key points. The subsequent 2D-3D coordinate transformation occurs using the depth information and camera intrinsics, where the RGB-D sensor is modeled as a pinhole camera as in (Pascual-Hernández et al., 2022). Each trajectory is represented by the Z-axis position of the pelvis joint, which serves as the root joint within the 24-joint SMPL body model. Additionally, a reference path—constructed and labeled through physical measurements—was included to provide a ground-truth baseline for comparison. By visualizing all three trajectories, we observed a strong alignment between the proposed method, YOLOv11, and the reference path. This close agreement indicates that our method reliably estimates the human trajectory and performs comparably to YOLOv11 in this context.

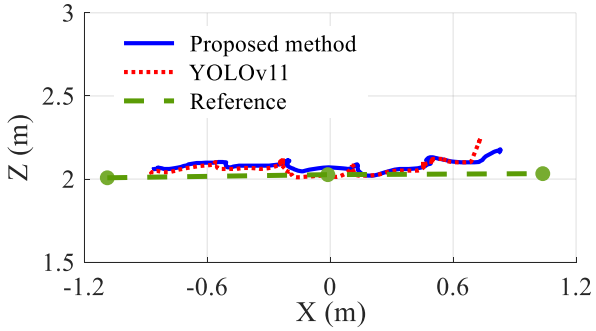


Figure 5. Pelvis motion trajectory comparison between the proposed method, YOLOv11, and the reference path.

Another key advantage of the proposed method is its robustness to noisy depth data and imperfect neural network predictions, as shown in Figure 6. The dashed line represents the estimated depth (distance along the camera's optical Z-axis) of the right-hand joint as predicted by YOLOv11. This depth estimate exhibits noticeable fluctuations and significant deviations from the expected value of ~ 2.0 m, as supported by Figure 5. In contrast, the same joint estimated using our proposed method produces a much smoother and more reliable depth trend, as illustrated by the solid line.

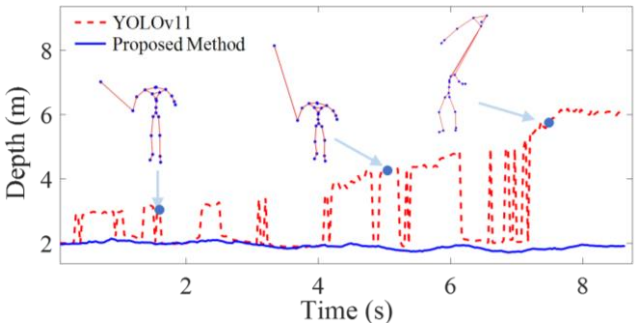


Figure 6. Comparison of right-hand depth estimates using the proposed method and YOLOv11. YOLOv11 results show significant errors at frames 45, 150, and 250 due to inaccurate right-hand identification.

The 3D body joints predicted by YOLOv11 at frames 45, 150, and 250, as shown in Figure 6, reveal larger estimation errors in the location of the right hand. In comparison, the results in Figure **7**, generated using our proposed method, demonstrate a significant improvement in estimation robustness. This is achieved by rescaling and positioning the pre-computed 3D avatar $M(\beta, \theta)$ within the global coordinate frame, effectively eliminating errors associated with individual body part identification.

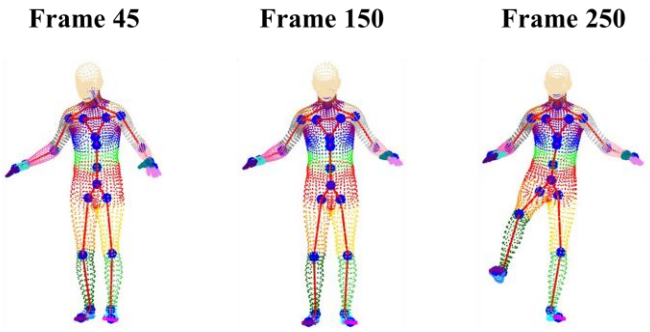


Figure 7. The estimated 3D human figure achieved by the proposed method. Significant improvement is observed compared to the YOLOv11 estimation in Figure 6.

4.2. Case Study 2: Human Pose Estimation with Robot

The second case study focuses on integrating the human and robot into a shared virtual environment to estimate their distance during workspace interaction. As shown in Figure 9, a single RGB-D camera monitors the HRC environment which consists of the human, reference path, and a Neuromeka Indyrp2v2 cobot arm. A digital twin of the robot is generated by streaming real-time joint angles from the robot controller and visualized in RViz, the Robot Operating System's 3D visualization tool. This setup enables synchronized analysis of both human and robot digital twins within a unified 3D space.

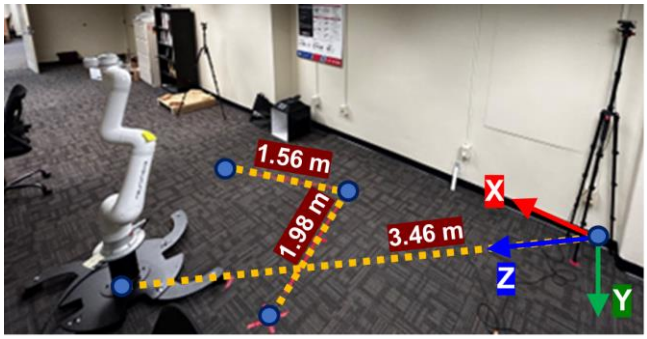


Figure 8. Experimental setup in case study 2.

Like the first case study, this experiment involves a human walking along a predefined path. However, in this scenario, the path is positioned in close proximity to an active robot arm, including segments where the human walks directly in front of the robot, causing partial visual occlusion from the camera's perspective. Consistent with the previous setup, data was collected using a single, fixed RGB-D camera. The human avatar is reconstructed using the proposed method and consists of 6890 surface points distributed across the full body, while the robot is similarly represented by a mesh of 566 surface points. These dense surface representations facilitate precise computation of the minimum distance between the human and the robot at each time step. By continuously monitoring these distances, the system can be extended to include real-time alerts for collision avoidance, thereby enhancing the safety of the shared workspace.

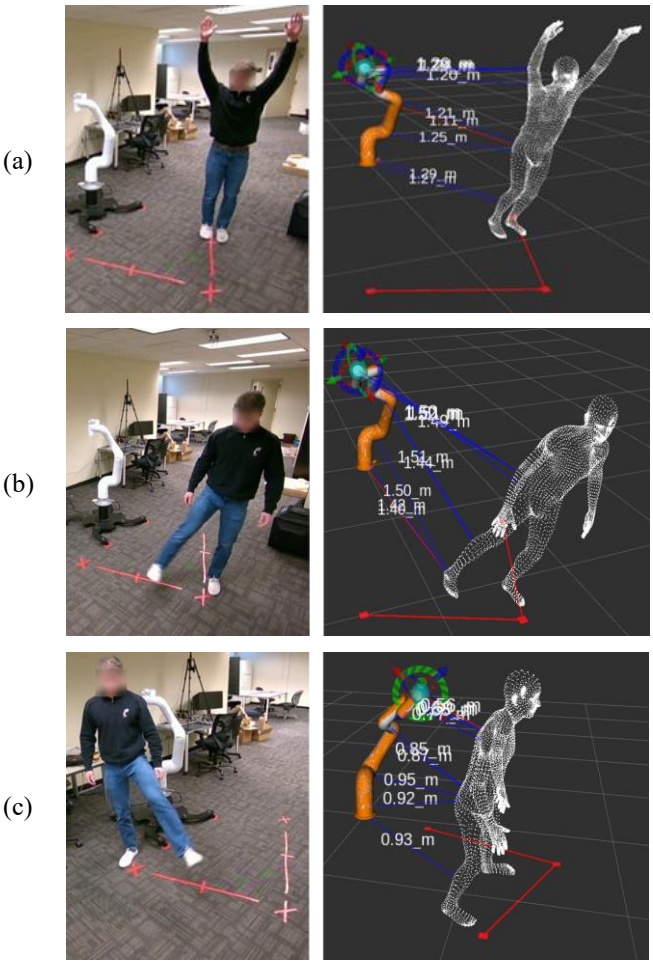


Figure 9. The estimated human-robot distance at different time frames (a) t=2.0s (b) t=4.13s (c) t=9.57s. Estimated distances between the human and select robot points are represented by blue lines, with the minimum distance represented by a red line.

The estimated distances between the human and robot across different time frames are represented in Figure 10. In this study, we do not utilize an expensive motion capture system to obtain ground-truth distances between human body joints and the robot. Addressing this limitation is part of our future work. Nevertheless, the results shown in the right column of Figure 9 clearly illustrate that the estimated lines connecting each body joint to the robot surface are reasonable and consistent with the expected spatial relationships, considering the walking path analysis in Figure 11. Although the robot is occluded from the camera's view in Figure 9(c), the proposed method is still able to reliably estimate the distance between the human and the robot. To the authors' knowledge, there is no existing backbone network capable of precise 3D body-part reconstruction. Therefore, the proposed method is currently the only known solution in literature capable of achieving accurate human-robot mesh estimation with localization.

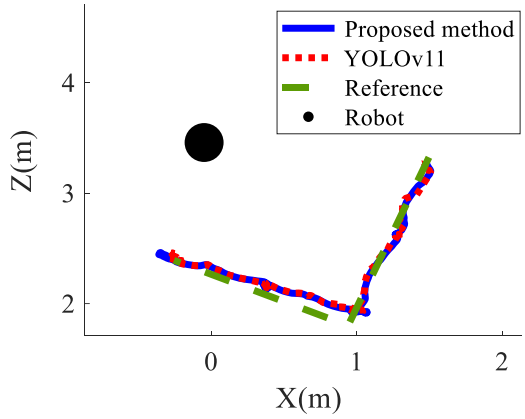
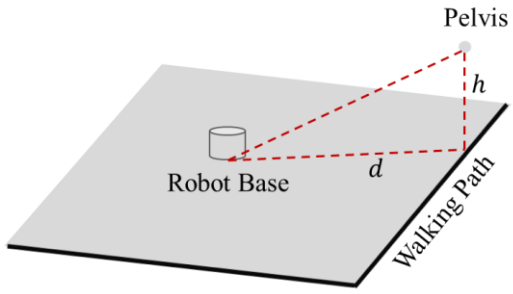


Figure 10. Pelvis motion trajectory comparison between the proposed method, YOLOv11, and the reference path.

(a)



(b)

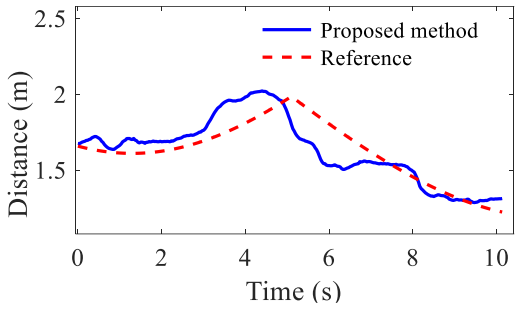


Figure 11 (a) an illustration of creating a reference distance between the human pelvis joint and the robot base. (b) comparison of the estimated distance using the proposed method against the reference.

The estimated human-robot distance is further validated, as shown in Figure 11, by comparing it to a reference measurement based on the lateral distance d from the robot base to the walking path and the human pelvis height h. In this analysis, h is assumed to be constant during movement and is measured prior to the experiment. The results shown in Figure 11(b) demonstrate that the distance estimates from the proposed method closely correspond to the calculated reference values. Building on the success of this preliminary experiment, we plan to quantitatively evaluate the estimation error using a calibrated motion capture system. This analysis will be addressed in future work.

In terms of latency, the framework achieves real-time performance in the monitoring loop: the robot's digital twin and human avatar are continuously updated in RViz at 30 FPS. The primary source of computational overhead lies in the mesh reconstruction stage. Ongoing work focuses on optimizing the reconstruction stage and exploring lightweight vision backbones to further reduce latency while preserving the accuracy of human shape and pose estimation.

5. CONCLUSIONS

This paper presents a novel framework for 3D sensing in HRC work cells, focused on real-time safety monitoring. The system uses input from a single RGB-D camera, processed through a fully neural network-based architecture, followed by depth-informed geometric transformations to reconstruct a detailed 3D human mesh in global coordinates. This mesh is integrated into a virtual environment alongside a real-time robot model, updated using data from the robot's controller. The setup enables continuous monitoring of human-robot interactions, including real-time distance measurement. Key contributions include: (1) accurate 3D reconstruction using depth data, (2) full mesh representation rather than a basic skeleton for improved safety, and (3) a neural network pipeline capable of real-time GPU deployment. Experimental results confirm the system's effectiveness, with precise 2D tracking of human walking paths and accurate 3D distance monitoring between the

human and robot, demonstrating the practical viability and reliability of the proposed method.

REFERENCES

Aivaliotis, P., Aivaliotis, S., Gkournelos, C., Kokkalis, K., Michalos, G., & Makris, S. (2019). Power and force limiting on industrial robots for human-robot collaboration. *Robotics and Computer-Integrated Manufacturing*, *59*, 346-360.

Byner, C., Matthias, B., & Ding, H. (2019). Dynamic speed and separation monitoring for collaborative robot applications–concepts and performance. *Robotics and Computer-Integrated Manufacturing*, *58*, 239-252.

Campomaggiore, A., Costanzo, M., Lettera, G., & Natale, C. (2019). A fuzzy inference approach to control robot speed in human-robot shared workspaces. ICINCO 2019-Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics,

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, *43*(1), 172-186.

Fang, H.-S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.-L., & Lu, C. (2022). Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE transactions on pattern analysis and machine intelligence*, *45*(6), 7157-7173.

Hirzinger, G., Butterfass, J., Fischer, M., Grebenstein, M., Hahnle, M., Liu, H., Schaefer, I., & Sporer, N. (2000). A mechatronics approach to the design of light-weight arms and multifingered hands. Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065),

IFR. (2023). *World Robotics 2023 Report: Asia ahead of Europe and the Americas*. https://ifr.org/ifr-press-releases/news/world-robotics-2023-report-asia-ahead-of-europe-and-the-americas

ISO, T. (2013). Robots and robotics devices-Safety requirements for industrial robots-Collaborative operation. In: Standard. Geneva, CH: International Organization for Standardization,(cited ….

Jiang, B. C., & Gainer Jr, C. A. (1987). A cause-and-effect analysis of robot accidents. *Journal of Occupational accidents*, *9*(1), 27-45.

Jocher, G., Qiu, J., & Chaurasia, A. (2024). Ultralytics YOLO11, Version 11.0. 0. In: Ultralytics.

Kamezaki, M., Wada, T., & Sugano, S. (2024). Dynamic collaborative workspace based on human interference estimation for safe and productive human-robot collaboration. *IEEE Robotics and Automation Letters*, *9*(7), 6568-6575.

Kocabas, M., Athanasiou, N., & Black, M. J. (2020). Vibe: Video inference for human body pose and shape estimation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

Lee, K., Shin, J., & Lim, J.-Y. (2021). Critical hazard factors in the risk assessments of industrial robots: causal analysis and case studies. *Safety and health at work*, *12*(4), 496-504.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2023). SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2* (pp. 851-866).

Magrini, E., Flacco, F., & De Luca, A. (2015). Control of generalized contact motion and force in physical human-robot interaction. 2015 IEEE international conference on robotics and automation (ICRA),

Marvel, J. A., & Norcross, R. (2017). Implementing speed and separation monitoring in collaborative robot workcells. *Robotics and Computer-Integrated Manufacturing*, *44*, 144-155.

Mohammed, A., Schmidt, B., & Wang, L. (2017). Active collision avoidance for human–robot collaboration driven by vision sensors. *International Journal of Computer Integrated Manufacturing*, *30*(9), 970-980.

NIOSH. (2022). *Robotics*. https://www.cdc.gov/niosh/topics/robotics/aboutthecenter.html

Pascual-Hernández, D., de Frutos, N. O., Mora-Jiménez, I., & Cañas-Plaza, J. M. (2022). Efficient 3D human pose estimation from RGBD sensors. *Displays*, *74*, 102225.

Patalas-Maliszewska, J., Dudek, A., Pajak, G., & Pajak, I. (2024). Working toward solving safety issues in human–robot collaboration: a case study for recognising collisions using machine learning algorithms. *Electronics*, *13*(4), 731.

Robla-Gómez, S., Becerra, V. M., Llata, J. R., Gonzalez-Sarabia, E., Torre-Ferrero, C., & Perez-Oria, J. (2017). Working together: A review on safe human-robot collaboration in industrial environments. *Ieee Access*, *5*, 26754-26773.

Shimada, S., Golyanik, V., Xu, W., & Theobalt, C. (2020). Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics (TOG)*, *39*(6), 1-16.

Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

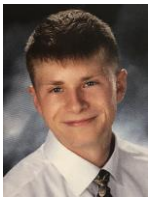
Zhang, Y., Kephart, J. O., & Ji, Q. (2024). Incorporating physics principles for precise human motion

prediction. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,

Zhang, Y., Xia, S., Chu, L., Yang, J., Wu, Q., & Pei, L. (2024). Dynamic inertial poser (dynaip): Part-based motion dynamics learning for enhanced human pose estimation with sparse inertial sensors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

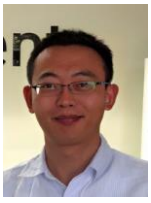
BIOGRAPHIES

Tarek Yahia is a Ph.D. Mechanical Engineering student at University of Cincinnati, Cincinnati, OH. He received his B.S. degree in Industrial Engineering from University of South Florida, Tampa, FL in 2024. His research interests include computer vision, image & signal processing, machine learning, and human-robot collaboration.



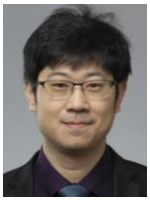
Kody Haubeil received his B.S. degree in mechanical engineering from Otterbein University, Westerville, OH, in 2024. He is currently pursuing his M.S. and Ph.D. degrees in mechanical engineering at the University of Cincinnati. His research interests include computer vision, machine learning, industrial A.I., and human-robot collaboration.

Alexander Suer received his B.S. and M.S. in mechanical engineering from the University of Cincinnati, Cincinnati, OH, USA. He is pursuing his Ph.D. degree in mechanical engineering at the University of Cincinnati. His research includes computer vision, language processing, physics informed neural networks, semiconductor PHM, and signal processing.



Yongzhi Qu is currently an assistant professor with the University of Utah. He earned his Ph.D. from the University of Illinois Chicago in 2014. His main research interests are machine learning for manufacturing automation, process and performance modeling, equipment diagnostics and prognostics.

Janet Dong is a professor in the Department of Mechanical and Materials Engineering and the directory of the UC Center for Robotics Research. She earned her Ph.D. degree in Mechanical Engineering from Columbia University, New York City, New York. Her research interests include robotics, manufacturing automation, dynamic systems and control, autonomous vehicles, mobile robots, industry 4.0/5.0, and product design and development.



Xiaodong Jia received the B.S. degree in engineering thermo-dynamics from Central South University, Changsha, China, in 2008, the M.S. degree in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2014, and the Ph.D. degree in mechanical engineering from the University of Cincinnati, Cincinnati, OH, USA, in 2018. He is currently an Assistant Professor with the Department of Mechanical and Materials Engineering, University of Cincinnati. His research interests include prognostics and health management, data mining, and ML.