2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

A Complementary Approach for Robust and Safety-Oriented Visual Tracking via Near-Infrared and RGB-D Cameras for Safe Physical Human-Robot Interaction

Mazin Hamad[1†], Samuel Kangwagye[1,2], Valentin Le Mesle[1], Rafael Mosberger[3,4], Achim J. Lilienthal[1,4], and Sami Haddadin[1,5]

***Abstract*— Safe physical human-robot interaction (pHRI) in industrial settings requires robust and accurate tracking of key points on the human co-worker's body and the robot structure. However, vision-based single-camera tracking solutions often face many challenges, such as limited field of view (FoV), detection range, occlusions, and inconsistent detection. This paper proposes a complementary multi-sensor tracking scheme that integrates RGB-D and near-infrared (NIR) cameras to improve human motion tracking accuracy while ensuring compliance with ISO/TS 15066 safety requirements. For the first time in pHRI, we deploy an infrared-based tracking system, originally designed for driver assistance and accident prevention, to complement RGB-D cameras, which provide detailed pose estimation at near range but suffer from a narrow FoV. A safety-oriented complementary approach is developed to fuse human tracking data from both systems into robot control, integrating a well-established safety paradigm based on the Safe Motion Unit (SMU) framework. The proposed system is experimentally validated in real-world collaborative robotic workspaces across various pHRI scenarios. Results demonstrate its effectiveness in respecting human safety constraints, even under challenging operating conditions, without unnecessary performance restrictions. The complementary vision-based approach improves tracking accuracy, expands FoV, and enhances reliability, making it a promising solution for certifiable, human-aware collaborative robotics in various industrial settings. The video documentation can be seen at `https://youtu.be/xWksc_vhuew`.**

I. Introduction

The manufacturing industry has undergone a profound transformation in recent decades, driven by the integration of robotic systems and smart sensors. This shift is paving the way for the so-called Fifth Industrial Revolution (in short: "Industry 5.0") [1], which places human well-being at the core of production processes [2], [3]. In this paradigm, advanced technologies such as robotic solutions are leveraged to augment workers' skills, taking over routine and cognitively less demanding operations while enabling them to focus on tasks that require creativity and complex decision-making. In this regard, collaborative robots–designed with lightweight, compliant structures [4] and equipped with collision detection and reaction strategies [5]–enhance process flexibility and task performance in terms of speed, precision, and payload capacity [6]. In future interactive manufacturing workcells, these collaborative robots are expected to work in close physical and cognitive collaboration with humans, forming synergistic teams that combine respective strengths to accomplish complex tasks [7]. However, ensuring human safety remains a primary challenge for fenceless robot operation, especially during physical human-robot interaction (pHRI) phases. To address this, certified sensor-enabled safety systems are crucial in detecting and mitigating potential hazards [8]. These systems typically integrate various sensor modalities to monitor the workspace, assess risks, and implement real-time safety measures, thereby protecting human co-workers and surrounding objects [9], see Fig. 1.

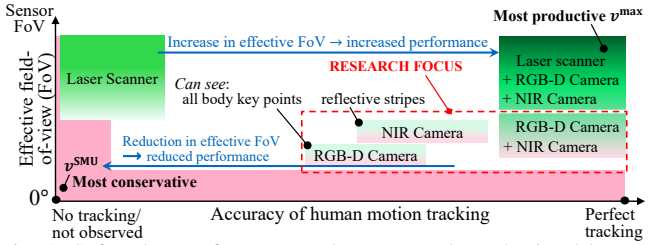


Fig. 1: Safe robot performance enhancement through visual human motion tracking: Increasing effective FoV and motion tracking accuracy enables a transition towards using a more productive speed.

The current robot safety regulations demand sensor redundancy and effective safety functions considering worst-case assumptions in safety-critical applications [10]. A list of the most relevant standards dealing with safety requirements in pHRI is provided in [11]. Recent robot safety standards, such as ISO/TS 15066 [12], specify requirements and guides for safe interaction and collaborative operation in industrial settings. For high task performance under safety constraints, typically achieved through efficient switching of the robot functional mode [13], precise and reliable tracking of key points on the human body and the robot structure is essential [14].

Motion tracking challenges for safe collaborative robots

To prevent accidents when human co-workers are collaborating with robots in shared workspaces and minimize injury risk when a robot-human collision is unavoidable, the real-time location of humans, ideally including positions and movement speeds of their limbs relative to the robot, must be reliably tracked in real time [15]. In dynamic, largely unpredictable environments, exteroceptive sensors are usually employed for this purpose [16]. In such scenarios, safety analysis must be thorough, enabling the evaluation of various system configurations and design modifications. However, as pHRI applications grow in complexity, certain risks may

This work was partly supported by the Federal Ministry of Research, Technology and Space of Germany in the programme of "Souverän. Digital. Vernetzt." Joint project 6G-life, project identification no. 16KISK002, by the Lighthouse Initiative Geriatronics by LongLeif GaPa gGmbH (Project Y), and by the European Union's Horizon 2020 research and innovation programme as part of the project DARKO under grant no. 101017274.

[1] Munich Institute of Robotics and Machine Intelligence (MIRMI), Technical University of Munich (TUM), Munich, Germany.
[2] Robotics and Automation Group, Department of Materials and Production, Aalborg University, Aalborg, Denmark.
[3] Retenua AB, Örebro, Sweden.
[4] Mobile Robotics and Olfaction Lab, Center for Applied Autonomous Sensor Systems, Örebro University, Sweden.
[5] Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), Masdar City, Abu Dhabi, United Arab Emirates.

[†] Corresponding author. Email: `mazin.hamad@tum.de`

be overlooked, especially those emerging over time. This limitation often leads to ineffective or overly conservative safety measures, such as maintaining unnecessarily low operational speeds [17]. The available vision-based solutions for safe pHRI rely on, at least, a measure of the distance between the robot and the human. Due to the inevitable inaccuracies introduced by intrinsic and extrinsic calibration for visual sensing, even with multi-sensory auto-calibration schemes [18], a safety margin must be added to the measured distance when the tracked human is physically interacting with the robot (cf. Fig. 1). Even though such a workaround might be enough e. g., for workspace monitoring or navigation tasks, efficient execution of complex collaborative manipulation tasks requires tighter interaction spaces while necessitating an accurate interplay of the used robot(s) with the involved human(s). Addressing these conflicting requirements while simultaneously ensuring human safety during pHRI emerges as a challenging research problem. Furthermore, ensuring safety certification compliance is another key challenge in pHRI, particularly in Industry 5.0 scenarios where humans and robots share workspaces and physically collaborate.

Research problem

To reduce the risk of common-cause failures in safety-critical systems, the *principle of diverse redundancy* is widely applied [19], [20]. For a collaborative robotic system to be safety certification-ready, it must incorporate redundant and independent sensing pipelines to meet functional safety requirements for collaborative operation alongside and with humans [21]. For this, a dual-camera tracking setup enhances the reliability of safe robot operation. It further ensures that the visual tracking system is not only robust in practice but also aligned with regulatory expectations from bodies such as, e.g., TÜV[1] in Europe. Utilizing this information to monitor the presence of humans in the vicinity and further track their motions relative to the robot, any well-established framework for safe pHRI, such as, e.g., the Safety Motion Unit (SMU) [22], can be utilized. To ensure safety without unnecessarily compromising performance, the safety system-based on the SMU framework in our case must address several tracking-related challenges. The key challenges include (cf. Fig. 1)

- **Limited sensor coverage/narrow field of view (FoV):** Single-camera solutions or setups with poorly located cameras struggle with near range tracking and may even lead to a loss of tracking, forcing unnecessary slowdowns.
- **Failures due to occlusions or poor detection:** Inconsistent tracking leads to overly conservative velocity scaling (hence, task performance reductions).
- **Lack of multi-sensor redundancy:** Redundancy is indispensable for safety-critical systems, making combining independent sensing pipelines essential for functional safety certification.
- **Unverified safety frameworks:** Existing tracking solutions are not yet fully integrated into validated, certification-ready architectures.

Considering visual sensors, the first three points result in the inability to detect or sufficiently track the human key points. Accordingly, the SMU enforces a conservative worst-case speed limit for the robot operation ($v^{\mathrm{SMU}} < v^{\mathrm{max}}$), reducing the task performance. Conversely, wider effective FoV with more accurate human tracking leads to improved task performance under safety constraints. In the ideal case of perfect tracking and sufficiently wide effective FoV, the robot can be operated even at the full desired productive speed ($v^{\mathrm{SMU}} = v^{\mathrm{max}}$), in case collisions are avoidable or pose no injury threats. Addressing the intersection of quality and accuracy of human motion tracking with safety issues contributes to delivering certifiable tracking solutions and safety systems for collaborative robots.

Contribution

This paper proposes a complementary vision-based multi-sensor approach to enhance safe human-robot collaboration. The proposed approach facilitates improving motion tracking accuracy and expanding the effective field of view, see Fig. 1, while supporting compliance with functional safety standards. Our scientific contributions can be summarized as follows:

I. We propose a novel complementary approach for robust and safety-oriented human motion tracking, that integrates RGB-D[2] and near-infrared (NIR) cameras to improve the accuracy, robustness, and reliability of human pose tracking in dynamic pHRI environments.
II. Integrating independent vision-based pipelines to leverage their complementary sensing modalities and strengths, the proposed approach is reliable and adheres to the functional safety certification requirements of having redundant components for safety-critical applications.
III. For validation, we test our approach empirically against relevant safety standards and evaluate the resulting human motion tracking accuracy, cycle time, and detection coverage, demonstrating compliance with ISO/TS 15066 and supporting the development of safety-certified human-aware collaborative robotic solutions.

The remainder of this paper is organized as follows. Section II presents the conceptual workspace sharing scenario and discusses the proposed setup of vision systems. Section III discusses the proposed safety-oriented complementary human motion tracking scheme. The experimental protocol is detailed in Section IV, where also the developed system setup is summarized, and the evaluation experiments are introduced, together with the evaluation metrics and the proposed safety-oriented visual sensor system. Section V provides the experimental results for exemplary use cases in a shared and collaborative human-robot environment. Finally, a comparative summary of safety-relevant features of different tracking systems under consideration is given. Section VI concludes the paper by highlighting the achievements and some future research directions.

II. Conceptual Workspace Sharing Scenario with Two Complimentary Vision Systems

We consider the conceptual workspace sharing scenario presented in Fig. 2, showing several human co-workers coexisting simultaneously and collaborating with different types of robots. For such a scenario, an essential visual perception requirement, from a safe pHRI perspective, is to provide robust and reliable human detection and motion tracking. For this, two vision-based systems are deployed

[1]TÜV is the short acronym for "Technischer Ãœberwachungsverein" in German, which means Technical Inspection Association in English.

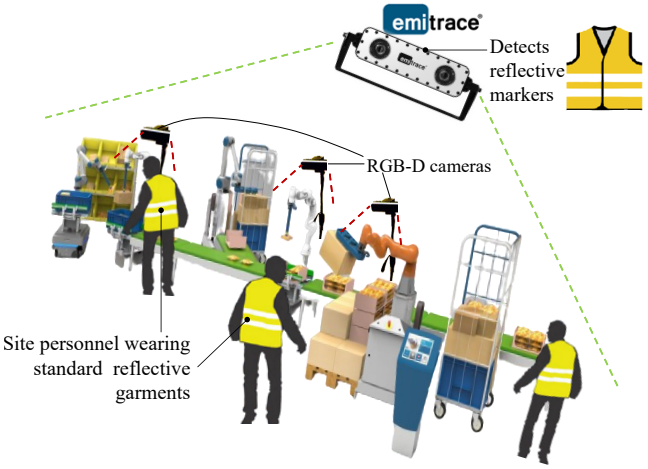
[2]RGB-D stands for "Red-Green-Blue and Depth". It refers to a type of imaging system that captures both color (RGB) information and depth (D) data of a scene, facilitating the 3D reconstruction of the real world [23].

to track the presence and movements of humans relative to the robot structures. We propose to monitor a large shared workspace area with an Emitrace camera, while multiple RGB-D cameras can be used for near range pose tracking of nearby human co-workers around each individual robotic workcell, as illustrated in Fig. 2(A). To differentiate the current workspace sharing mode (coexistence vs. close collaboration) [13], [24], we propose to define dynamic interaction regions of *coexistence* and *collaboration* as shown in Fig. 2(B) for the case of a single robotic workcell. Their instantaneous borders are defined based on the current locations of both the robot's points-of-interest (POIs) and the involved human key points.

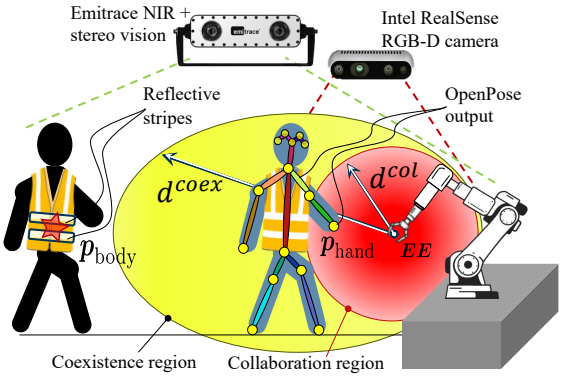
OpenPose/RGB-D based system for human pose tracking: The OpenPose human pose tracking setup employs an RGB-D camera with a narrow field-of-view (FoV) for near range pose tracking of nearby human co-workers around each robotic workcell. It is also integrated with a vision-based tracking system that estimates the position of body key points of a human co-worker relative to the robot [14]. As shown in Fig. 2(B), the human key points are obtained from a vision processing pipeline that uses an Intel RealSense D435 RGB-D camera to capture the 2D images and OpenPose with the BODY_25 model [25] for the identification of human key points in these images. The key points are then de-projected using the aligned depth image from the RealSense Python API to obtain the position of human key points in 3D Cartesian space with the reference frame origin at the robot base.

Emitrace system for tracking reflective garments: As a secondary vision system, we propose deploying an Emitrace [26] system and extending its tracking capabilities for collaborative human-robot applications. By well-positioning this near-infrared (NIR) based stereo camera, one can achieve a large 3D FoV that enables reliable human presence detection by identifying standard reflective garments, as illustrated in Fig. 2 (right side). A novel, robust camera-based system for detecting humans wearing safety clothing with reflective markers was first described in [27]. The system's latest version combines the advantages of NIR stereo and a monocular color camera vision [28]. This combination allows efficient detection of reflective patterns, inferring their distance to the camera, and generating regions of interest (ROIs) for 3D tracking and pose estimation of people wearing safety garments [29]. A commercial, fully embedded vision-based system for intelligent driver assistance and accident prevention on heavy industrial vehicles, termed Emitrace, is available today [26]. The Emitrace tracking system was extensively evaluated in many indoor and challenging outdoor scenarios, and it showed good classification performance and accurate distance estimation. However, it has not been yet applied in usage scenarios of collaborative robotics such as the one shown in Fig. 2.

To summarize, the OpenPose setup employs a relatively inexpensive RGB-D camera to estimate the human skeleton and pose. However, it only works at close distances (i.e., for near range monitoring and tracking), is less accurate (due to depth estimation with no active sensors), and has a slow update cycle (as a result of its visual information processing pipeline). As shown in Fig. 2(B), the RGB-D camera can cover the collaboration region and very little of the coexistence region. On the other hand, the Emitrace system has a wider FoV and covers both collaboration and coexistence regions. It can detect reflective markers even



(A) Conceptual large workspace with one Emitrace camera for monitoring co-workers in the whole workspace and several robotic workcells (each with one RGB-D camera at least for near range motion tracking).



(B) One robotic workcell extracted from (A) showing interaction regions. p_{body} and p_{hand} are body and hand positions, while d^{col} and d^{coex} are threshold distances for collaboration and coexistence ranges, respectively.

Fig. 2: A conceptual human-robot workspace sharing scenario showing one Emitrace camera and several robotic workcells, each with RGB-D cameras. The RGB-D camera tracks the co-worker's body parts by pose estimation, while Emitrace tracks the reflective markers on reflective garments worn by humans.

from far distances across a long range throughout the whole dynamic workspace under changing lighting conditions via active NIR sensing. However, possessing these great tracking capabilities comes at more expensive Emitrace system price and relatively higher installation costs. Moreover, the commercial Emitrace system has no pose tracking capability and it has not been implemented on collaborative robots before. A unified system that combines the tracking outputs of these two vision systems in a complementary fashion within an established safety paradigm employing the SMU framework to gain most of their advantages and avoid their shortcomings is developed in the following.

III. Safety-Oriented Complementary Human Tracking Scheme

The schematic description of our proposed perception-safety scheme, based on the complementary, multi-camera tracking system, is presented in Fig. 3. The proposed scheme defines and employs dynamic interaction regions to simultaneously and optimally satisfy the ISO/TS 15066 requirements for collaborative operation [12] under different interaction scenarios. The detailed setup of the cameras and the robot in Fig. 3 is depicted in Fig. 2. The rest of the parts

are discussed in detail hereafter.

A. Information Processing

As shown in Fig. 3, the set of human(s) motion tracking data from the cameras is sent and received at the robot side over a Remote Procedure Control (RPC). A Kalman filter (KF) is then applied to the received tracking outputs in order to obtain smoother motion estimates. The KF uses a constant velocity process model to simulate human motion, as recently recommended in [30]. Its outputs are smoothened position and estimated velocity. For all received human key point positions, the distance to the robot POI[3] is calculated. The closest distance h_{dist} is chosen to determine if the human (e. g., hand) is in collaboration ($h_{dist} \leq d^{col}$) or the human body (e. g., chest) is in coexistence range ($d^{col} \leq h_{dist} \leq d^{coex}$). To this end, the calculated relative distances are checked, and assigned thresholds of coexistence and collaboration speeds are applied to limit the robot's EE task speed. The tracked position data of both cameras is sent to the robot, where the velocities are estimated using the Kalam filter. The relative human-robot velocities are monitored by the task/robot controller and, if necessary, filtered by the SMU to ensure human safety upon any potential, undesired collision. The fusion strategy employs a priority-based approach: RGB-D tracking takes precedence when available due to its full body posture tracking, with automatic fallback to NIR tracking of upper torso movement when RGB-D loses detection. A Kalman filter smooths transitions between sensors to prevent abrupt motion changes. While a multi-modal information fusion was considered [31], [32], the priority-based switching was chosen for its simplicity and deterministic safety behavior required for ISO/TS 15066 compliance. Occlusions are handled via complementary sensing, so when one sensor is occluded the system automatically switches to the alternate sensor with more realistic distance measurements.

B. Safe Motion Unit (SMU) Implementation

To enable assessing the effects of tracking quality on safety and task performance, the proposed complementary visual tracking scheme is investigated using the SMU [22] as the underlying safety framework. The primary function of the SMU framework is to ensure generating safe task velocity limits in case of potential, undesired collisions (hence, only safe robot motions w. r. t. to human injury biomechanics), $v^{\max}$. This is achieved based on the most conservative interaction region activated by the relative distance between humans and robots, i.e., coexistence, d^{coex} or collaboration, d^{col}. For our purposes in this work, the safe thresholds for d^{coex} and d^{col} are directly assigned threshold values based on biomechanical safety knowledge [33] since we focus rather on demonstrating the proposed complementary tracking scheme and assessing its performance under safety constraints, and not on the SMU implementation. The details of computing d^{coex} and d^{col} based on a unified manner to satisfy the ISO/TS 15066 safety requirements, in terms of Speed and Separation Monitoring (SSM) and Power and Force Limiting (PFL) [12], can be found in [14]. Furthermore, in correspondence to the described interaction regions, switching these functional operation modes of the collaborative robot is done as detailed in [13].

[3]For simplicity and without loss of generality, to clearly explain the developed concept, we assign and consider a sole POI at the robot end-effector (EE). Other POIs can be also assigned at any other point on the robot structure and with a minimal geometrical extension considered [22].

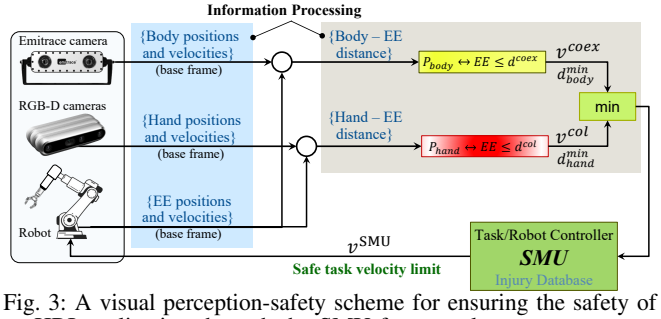


Fig. 3: A visual perception-safety scheme for ensuring the safety of a pHRI application through the SMU framework.

IV. Materials and Methods

A. System and Sensor Hardware Setup

All experiments described in this work were executed with a 7-DoF[4] industrial, lightweight robotics manipulator arm (namely, the Franka Emika Panda [34]). To track the motion of different human body parts in the close vicinity of the robot, the 2D images of RealSense were used within an OpenPose/RealSense based key points estimation algorithm (denoted OpenPose/RS in our results). Furthermore, both employed cameras (RealSense and Emitrace) were calibrated relative to the robot base. This was achieved by utilizing AruCo markers for RealSense and a calibration object with reflective markers for Emitrace. Image processing was done on a Media PC with a dedicated GPU and the Open Source Computer Vision (OpenCV) library v3.0 [35]. An RPC server/client protocol was used to send the tracking results from both vision systems to the robot task controller.

B. Ground truth 3D Motion Data from the Robot

The setup for obtaining ground truth information is illustrated in Fig. 4. A very accurate estimation of the 3D position, $\boldsymbol{p}(t)$, and velocity, $\boldsymbol{v}(t)$, of the end-effector[5] are recovered from the robot kinematics by means of internal sensing as follows

$$\begin{cases} \boldsymbol{p}(t) = \boldsymbol{t}(1:3)\,; \;\; {}^{\mathcal{E}}_{\mathcal{B}}\boldsymbol{T}(\boldsymbol{q}(t)) = \Big(\boldsymbol{R}(t) \;\big|\; \boldsymbol{t}(t)\Big) \\ \boldsymbol{v}(t) = \boldsymbol{J}_v(\boldsymbol{q}(t))\dot{\boldsymbol{q}}(t)\,; \;\; \boldsymbol{J}(\boldsymbol{q}(t)) = \begin{bmatrix} \boldsymbol{J}_v(\boldsymbol{q}(t)) \\ \boldsymbol{J}_\omega(\boldsymbol{q}(t)) \end{bmatrix} \end{cases}, \tag{1}$$

where $\boldsymbol{q}(t) \in \mathbb{R}^n$ denotes the robot generalized joint coordinates (i.e., its joint configuration) at time t, ${}^{\mathcal{E}}_{\mathcal{B}}\boldsymbol{T}(\boldsymbol{q}(t)) \in \mathbb{R}^{4\times4}$ denotes the robot homogeneous transformation matrix summarizing its forward kinematics from the base frame $\mathcal{B}$ to the end-effector frame $\mathcal{E}$ with $\boldsymbol{R}(t) \in \mathbb{R}^{3\times3}$ and $\boldsymbol{t}(t) \in \mathbb{R}^{4\times1}$ being the configuration-dependent rotation matrix and the extended homogeneous translation vector, respectively. The Jacobian matrix associated with the robot end-effector point EE is $\boldsymbol{J}(\boldsymbol{q}(t)) \in \mathbb{R}^{6\times n}$, with $\boldsymbol{J}_v(\boldsymbol{q}(t)) \in \mathbb{R}^{3\times n}$ and $\boldsymbol{J}_\omega(\boldsymbol{q}_r(t)) \in \mathbb{R}^{3\times n}$ denoting the sub-Jacobians for translational and rotational motions, respectively.

C. Metrics for Evaluating the Tracking Performance

The following metrics are adopted to evaluate the performance of the proposed complementary tracking scheme and compare it against the cases when Emitrace and RGB-D vision systems are used individually.

[4]DoF is short for *degrees of freedom*.

[5]The rotational motions are completely neglected, as estimating object orientation goes beyond the capabilities of the investigated vision systems and, therefore, it is not within the scope of this work.

- **Tracking accuracy** for quantifying the accuracy of the vision system in detecting/estimating the perceived object's location and further tracking its motion. This metric is evaluated by calculating the norm of 3D position estimation against the ground truth.
- **Tracking cycle time** captures the rate at which the visual perception pipeline outputs the tracked object's position data. An average of multiple runs involving long tracking cycles is used to quantify this metric at the receiver's side (i.e., the robot task controller).
- **Detection volume per field-of-view and range** indicates how many objects in the camera 3D FoV (horizontal and vertical) within its detection range can be correctly tracked simultaneously, e.g., without depth occlusion.
- Other **qualitative metrics**, which include, besides the nature of the provided separation distance (static vs. dynamic), that can be evaluated at the robot side, any ISO/TS15066 deduced safety requirements or functions according to the well-established taxonomy introduced in [36].

D. Investigated Scenarios

The proposed perception-safety scheme was tested on two human-robot workspace sharing and collaboration scenarios:

1) Within a shared workspace inside a structured laboratory environment with stable lighting conditions, where an SMU-controlled robot arm is employed to execute a pick-and-place task of moving beverage bottles from one box to another.
2) Simulating an unstructured, challenging industrial working environment, where complex situations such as changing lighting conditions may occur while multiple human co-workers wearing safety vests move around and possibly interact with the SMU-controlled robot.

In both scenarios, the human co-workers put on conventional, off-the-shelf safety vests, and we relied on the reflective marker pattern at the chest for Emitrace detections. Additionally, the closest body part of all human co-workers alternates between chest (tracked by Emitrace system) and hand (tracked by OpenPose/RGB-D system), which may interact with the robot end-effector while it is carrying out the pick-and-place collaborative task. For our investigated scenarios, the following thresholds were used within the integrated SMU framework. The maximum achievable task speed for the robot end-effector, $v^{\max}$, was set to 0.45 $\mathrm{m/s}$. We further assumed a coexistence, safe speed limit, v^{coex}, of 0.3 $\mathrm{m/s}$ and a more restrictive collaboration speed limit, v^{col}, of 0.15 $\mathrm{m/s}$. The coexistence distance threshold was set to 0.8 m, while that of the narrower collaboration region was set to 0.3 m (both with a hysteresis region of $\pm 5\%$).

E. Experiment I: Tracking Comparison and Insights

The goal of this experiment was to assess quantitatively the performance of both tracking systems in terms of 3D position estimation accuracy and measured cycle time. We aimed to evaluate the tracking characteristics of the two vision systems and further reason objectively about utilizing their tracking outputs in a complementary fashion for safety-oriented tracking of human co-workers in shared, pHRI-enabled working environments.

Since the OpenPose/RGB-D setup can only estimate the human skeleton and locations of its body segments, the experiment was designed such that both vision systems have to estimate the same wrist location. For this, a marker pattern comprised of two horizontal parallel reflective stripes, simulating a mini safety vest, was used to recover the hand motion by the Emitrace system. Its attachment location was chosen at the same location the OpenPose/RGB-D setup reports the wrist key point, see Fig. 4 (top, left). The human subject (operator/co-worker) grabbed firmly (at a marked grasping spot) a handle attached to the robot end-effector directly at point EE. The technical objective was, therefore, to estimate the end-effector position from the wrist location detected by cameras, by simply using a transformation matrix from the wrist location to EE. This way, the estimates for the end-effector motion can be obtained from the very accurate proprioceptive joint position-sensing of the robot itself. The retrieved 3D positions and the corresponding velocities (obtained via differentiations with good filtering inside the robot controller) were considered as our *ground truth* for later comparisons against different tracking outputs under consideration.

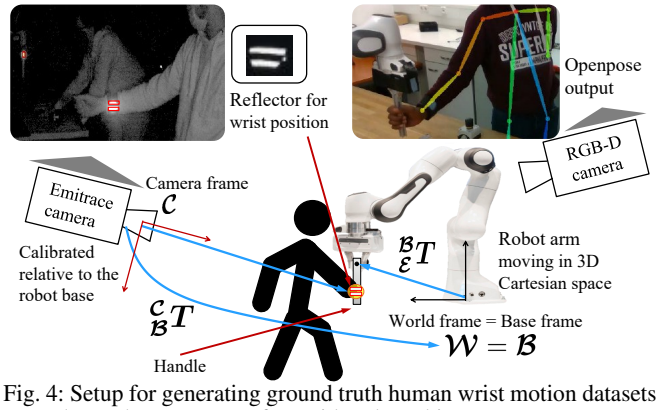


Fig. 4: Setup for generating ground truth human wrist motion datasets to evaluate the accuracy of considered tracking systems.

F. Experiment II: Complementary Tracking in Different Environments

As mentioned earlier, the Emitrace system has a wider FoV and detects reflective markers along a larger range of distances (along the depth space). On the other hand, the OpenPose setup employs an inexpensive RGB-D camera to estimate the human skeleton and pose, but its detection concept only works at close distances (i.e., in near ranges). Accordingly, in this experiment, human-robot collaborative scenarios were utilized to show how to use the tracking outputs of these two vision systems in a complementary fashion to gain most of their advantages and avoid their shortcomings.

V. Results and Discussion

A. Experiment I Results

By comparing the recovered 3D position of the robot end-effector with each of the corresponding individual camera measurements, position estimation error profiles of the two vision systems were estimated as shown in Fig. 6 (left) for a sample 3D triangular motion. As can be seen from the plots, both camera measurements are very noisy when examined within the much faster robot's control cycle (running at 1 kHz). Accordingly, any relative distance or velocity estimations using these measurements directly will also be noisy. Indeed, this is unacceptable from a safe robot control point of view, as it may lead to jerky robot motions and dangerous structural vibrations. As a remedy, a Kalman filter was utilized to

smoothen out the raw position measurements, see Fig. 6 (left). The target's velocity was also estimated with the added Kalman filter to provide more insights about the tracking quality. The corresponding smooth velocity estimates for the robot end-effector are plotted in Fig. 7 versus the ground truth retrieved from the robot proprioceptive joint-sensing.

Averaging over multiple runs for each of the two executed 3D motions (runs $1-6$ were linear while runs $7-9$ were triangular), the mean norm values of the 3D position errors for the two vision-based tracking systems were obtained as reported in Fig. 6 (right). From the shown bar graph, it is evident that the Emitrace based tracking system was able to track the designed wrist marker with an average position error of less than 1.4 cm versus 2.4 cm for the OpenPose/RealSense based tracking system. Applying the smoothing Kalman filter, with the same settings for both systems, further improves accuracy to less than 1.2 cm and 2.2 cm norm of position error for the Emitrace system and the OpenPose/RealSense system, respectively. The RGB-D system achieved a detection success rate of consistently above 95% at close range (<1.5m), while Emitrace maintained above 90% detection with reflective markers across the full workspace.

Regarding the tracking performance, the Emitrace setup required more than 63 ms to update the tracking data (as estimated on the receiver/robot side), compared to less than 27 ms for the OpenPose/RealSense setup. The average cycle time duration for each vision-based tracking system is listed in Table I, together with the nominal volume of detected objects per available 3D FoV and detection range.

B. Experiment II Results

For the shared laboratory scenario with stable lighting conditions, the collaborative task was repeated multiple times to allow the human co-workers to excite different patterns for moving between interaction regions during the task execution. The results in terms of relative distances between the robot end-effector POI and the two closest co-workers (tracked with OpenPose/RGB-D system), as well as commanded SMU task velocity limits, are shown in Fig. 8 (left). Various dynamically-defined regions for interaction and the chosen coexistence and collaboration velocity limits are also indicated. At the beginning of the motion, the robot could move with the full task speed $v^{\max}$ as fast as 0.45 $\mathrm{m/s}$ since the closest co-worker was away (in the monitored area); check the bottom plots. As soon as the co-worker entered the coexistence region (at $t = 5$ s) and stayed there for a while but did not further approach the collaboration region, the proposed perception-safety complementary scheme set the task velocity via the SMU to the less conservative limit. When the co-worker started collaborating with the robot using his hand (at $t = 9$ s), the SMU reduced the maximum allowable task velocity even further to the more conservative, prespecified collaboration limit. On the contrary, when the co-worker was not heavily collaborating with the robot (shortly after $t = 22$ s), the SMU relaxed the safety constraint on the task velocity again to the coexistence limit. This was kept until the co-worker moved outside the coexistence region (at $t = 26$ s), where the SMU recovered the full task velocity (i.e., $v^{\mathrm{SMU}} = v^{\max}$). These autonomous safety routines run indefinitely within the robot control cycle as long as it executes motion tasks in the shared workspace.

Similar physical interaction patterns to the previous scenario were repeated for the scenario simulating the unstructured environment with the lighting being switched off and on frequently. These were deliberately triggered at time instants $t = 7$, 17, 27, 37, and 42 s. The results for this scenario are shown in Fig. 8 (right). Accordingly, one can see that the OpenPose/RealSense based system lost track due to the resulting sudden darkness at these specific time instants after the lighting loss. However, thanks to the integrated NIR stereo of the Emitrace system and our perception-safety integration scheme, the robot always retreated to use the most conservative safe velocity limit from the Emitrace vest/chest tracker. This verifies the robustness of the proposed complementary tracking system under such challenging operating conditions. It further shows its reliable capabilities in ensuring the safety of human co-workers around the robot, without the need to completely stop its task execution or reduce its task performance more than what is actually necessary.

C. Comparative Summary

After gathering all the needed information, we qualitatively assessed the performance in terms of tracking cycle, detection accuracy, and detection volume per FoV per range. This comparative assessment was done for the cases when the Emitrace system and RGB-D systems were utilized individually against when the proposed complementary system was used. The experimental findings are summarized in Table I. The proposed complementary tracking approach outperforms the two individual ones in all criteria with the fastest tracking cycle of 0.027 s, the best detection accuracy of 0.0241 m, and the largest detection volume of Up to *scene size*$/(100^\circ \times 70^\circ) * 15\,\mathrm{m}$. Regarding the ISO/TS 15066 deduced requirements and safety functions, the proposed safety-oriented scheme combines the advantages of the two investigated vision-based systems and further transcends their individual capabilities. This aspect, together with the extended 3D FoV and detection range, offers new desired functionalities for safe pHRI (e. g., NIR vision for robust tracking of human co-workers in unstructured environments). The quantitative performance improvements are evident in the metrics presented in Table I and the detailed tracking accuracy analysis in Figs. 6-8. A summary video documenting technical concepts, all experimental settings and key results can be seen online at https://youtu.be/xWksc_vhuew.

VI. Conclusion

This paper advances safety in pHRI by developing a robust and reliable human detection and tracking scheme. We proposed a safety-oriented complementary approach that integrates tracking information from RGB-D and near-infrared cameras, leveraging their complementary strengths for enhanced human motion tracking accuracy and robustness. Our study marks the first application of the Emitrace NIR camera in pHRI for collaborative robotics,[6] demonstrating its effectiveness beyond its original application domain.

The tracking performance of the RGB-D and Emitrace systems was individually assessed using key evaluation metrics, including position tracking accuracy, cycle time, detection volume, and field-of-view coverage. Despite not being originally designed for robotic applications, the Emitrace system demonstrated sufficient tracking accuracy for the investigated pHRI scenarios. The proposed complementary

[6]The Emitrace tracking system was originally designed for driver assistance and accident prevention in industrial vehicles and mobile machinery [26].

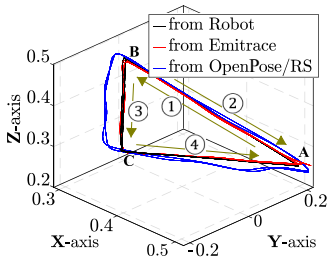


Fig. 5: Experiment I motion profiles: 1) linear path AB (① followed by ②), or 2) triangular path ABC (①, ③ then ④).

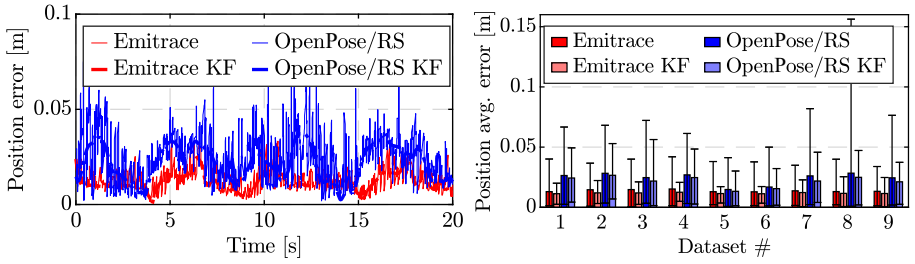


Fig. 6: Experiment I: Accuracy of the position estimation for the motions in Fig. 5. Position tracking accuracy was estimated for a triangular motion (shown on the left), where the bar graph summarizes the time average, minimum, and maximum of the tracking error norms for multiple runs (right).

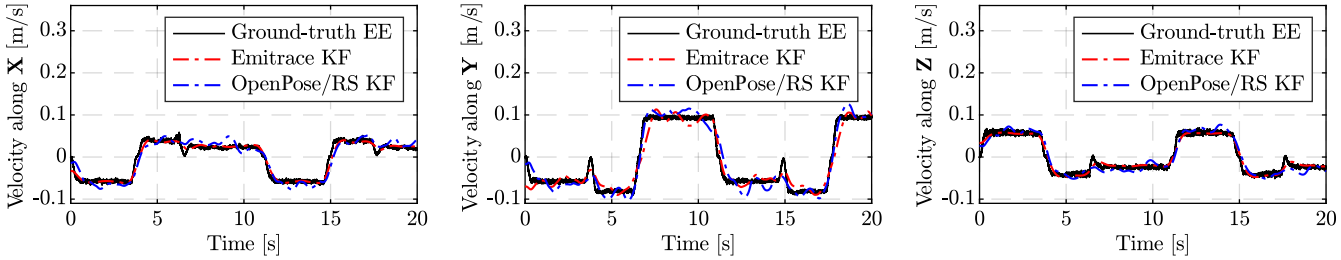


Fig. 7: Estimated 3D velocity profile for EE triangular motion (position error norm in Fig. 6, left) versus the ground truth.

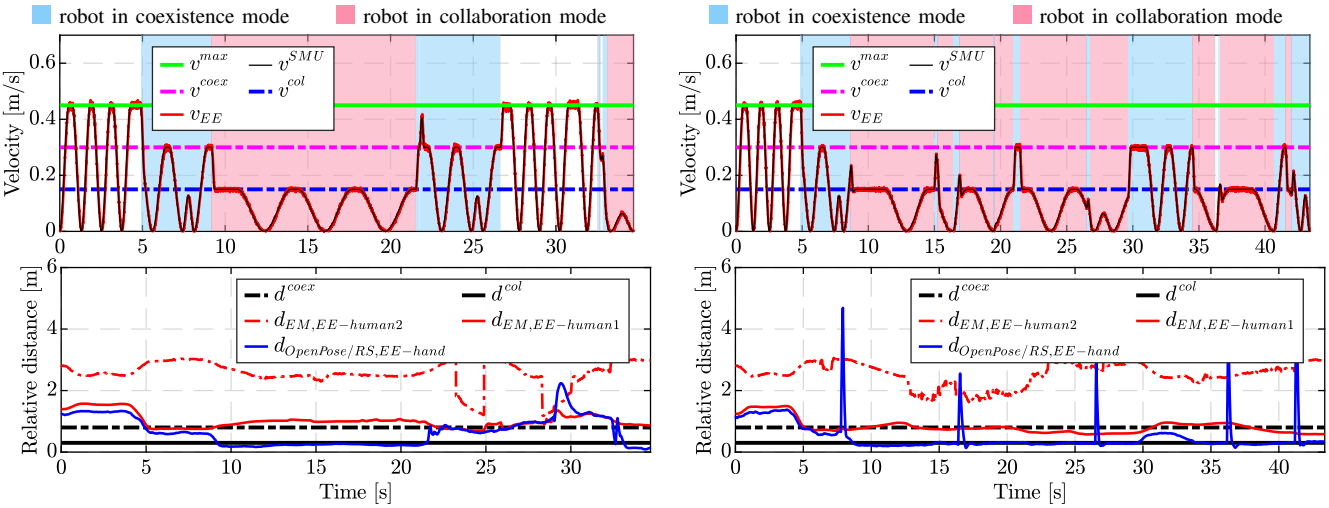


Fig. 8: Experiment II: Distances and velocities relative to the robot end-effector in different working environments. Distances to the hand of the closest co-worker (*human1*) and to the chests of the closest and second closest co-worker (*human2*) are shown at the bottom. In the upper plots, the active operational mode is color-coded. Runs in a controlled laboratory environment with stable lighting conditions are shown on the left, while lighting losses during the other runs are shown on the right as jumps in the OpenPose/RealSense based estimates.

TABLE I: Comparative summary of the investigated vision-based tracking systems for safe pHRI.

Vision system	**Used sensors**	**Cycle time*** **[s]**	**Detection accuracy** [m]	**Detection volume/ FoV*range** [(°×°) * m]	**ISO/TS15066 deduced safety requirements *or* safety function according to the taxonomy in [36]**	**Other notable features**
Emitrace [28] (Embedded AI-powered vision technology)	Hybrid RGB/NIR camera	0.063	0.0137	Up to scene size/ $(100°\times70°)*15\,\mathrm{m}$	Robot position control, robot speed control (based on relative human-robot motion, both distance and speed)	Industrial-grade AI-powered vision system, vision under challenging lighting conditions through NIR active sensing
OpenPose/RealSense human skeleton key points tracker [14]	RGB-D camera	0.027	0.0241	Up to scene size/ $(69°\times42°)*1.5\,\mathrm{m}$	Robot position control, robot speed control (based on the distance between the robot and the human), near field vision system: upper body function, hand function	Full-body human skeleton recovery, pose estimation and motion tracking
Complementary tracking system (*this work*)	Emitrace + OpenPose/ RGB-D combined	0.027	0.0241	Up to scene size/ $(100°\times70°)*15\,\mathrm{m}$	Safety-rated monitored stop, robot position control, robot speed control (based on relative human-robot motion), near-field tracking system for upper body and hands	Dynamic separation distance, full integration with the SMU framework, conformity with the ISO/TS 15066 SSM and PFL modes, robustness against challenging lighting conditions

*Cycle time is tracking mode dependent: 0.027 s when RGB-D is used *versus* 0.063 s when the tracking system falls back to NIR sensing only.

tracking system was then integrated into a well-established safety paradigm, ensuring human safety under various pHRI conditions. Experimental results verified that the system respects safety constraints even under changing lighting

conditions, without unnecessary performance restrictions.

By integrating two independent sensing pipelines, the proposed approach meets the minimum redundancy requirements for functional safety certification, paving the way toward certifiable collaborative robotic workcells. While the current work demonstrates effectiveness in controlled scenarios, scaling to larger industrial deployments with multiple workcells presents challenges in terms of camera placement optimization and computational resource allocation. Future research will further explore the benefits of complementary vision-based tracking, focusing on optimizing Emitrace integration with cost-effective RGB-D based tracking systems. Specifically, we will investigate its application in larger workspaces and broader industrial robotics scenarios with multiple robotic workcells and several human co-workers moving around and collaborating with robots. This will help further justify the cost-effectiveness of integrating Emitrace cameras for collaborative robotics while strengthening their role in safety certification efforts.

ACKNOWLEDGMENT

The authors thank Alexander Kurdas and Chenxing Li for their support and discussions about integrating the Emitrace camera to track humans in indoor settings.

Please note that both R. Mosberger and A. Lilienthal have a potential conflict of interest as shareholders of Retenua AB (Sweden); the startup behind the development of the Emitrace NIR tracking system.

REFERENCES

[1] T. S. Chu, A. B. Culaba, and J. A. C. Jose, "Robotics in the fifth industrial revolution," in *Intl. Conf. on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*. IEEE, 2022, pp. 1–6.

[2] S. Huang, B. Wang, X. Li, P. Zheng, D. Mourtzis, and L. Wang, "Industry 5.0 and Society 5.0 – Comparison, complementation and co-evolution," *Journal of Manufacturing Systems*, vol. 64, pp. 424–428, 2022.

[3] B. Vanderborght, "Unlocking the potential of industrial human-robot collaboration: A vision on industrial collaborative robots for economy and society," *Publications Office of the European Union*, 2020.

[4] A. Bicchi and G. Tonietti, "Fast and" soft-arm" tactics [robot arm design]," *IEEE Robotics & Automation Magazine*, vol. 11, no. 2, pp. 22–33, 2004.

[5] S. Haddadin, A. De Luca, and A. Albu-Schäffer, "Robot collisions: A survey on detection, isolation, and identification," *IEEE Transactions on Robotics (T-RO)*, vol. 33, no. 6, pp. 1292–1312, 2017.

[6] E. Colgate, A. Bicchi, M. A. Peshkin, and J. E. Colgate, "Safety for physical human-robot interaction," in *Springer Handbook of Robotics*. Springer, 2008, pp. 1335–1348.

[7] R. Ahmad and P. Plapper, "Human-robot collaboration: Twofold strategy algorithm to avoid collisions using tof sensor," *International Journal of Materials, Mechanics and Manufacturing*, vol. 4, no. 2, pp. 144–147, 2015.

[8] International Organization for Standardization (ISO), "ISO 12100:2010 – Safety of machinery – General principles for design – risk assessment and risk reduction," 2010.

[9] C. Scholz, H.-L. Cao, E. Imrith, N. Roshandel, H. Firouzipouyaei, A. Burkiewicz, M. Amighi, S. Menet, D. W. Sisavath, A. Paolillo *et al.*, "Sensor-enabled safety systems for human-robot collaboration: A review," *IEEE Sensors Journal*, 2024.

[10] J. Starr and C. Quick, *Robotic Safety Systems: An Applied Approach*. CRC Press, 2024.

[11] M. Valori, A. Scibilia, I. Fassi, J. Saenz, R. Behrens, S. Herbster, C. Bidard, E. Lucet, A. Magisson, L. Schaake *et al.*, "Validating safety in human–robot collaboration: Standards and new perspectives," *Robotics*, vol. 10, no. 2, p. 65, 2021.

[12] International Organization for Standardization (ISO), "ISO/TS 15066 Robots and robotic devices – Collaborative robots," 2016.

[13] P. Svarny, M. Hamad, A. Kurdas, M. Hoffmann, S. Abdolshah, and S. Haddadin, "Functional mode switching for safe and efficient human-robot interaction," in *IEEE-RAS Intl. Conf. on Humanoid Robots (Humanoids)*, 2022, pp. 888–894.

[14] P. Svarny, M. Tesar, J. K. Behrens, and M. Hoffmann, "Safe physical HRI: Toward a unified treatment of speed and separation monitoring together with power and force limiting," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2019, pp. 7580–7587.

[15] 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.

[16] M. Ben-Ari and F. Mondada, "Sensors," in *Elements of Robotics*. Springer, 2018, pp. 21–37.

[17] F. Vicentini, M. Askarpour, M. G. Rossi, and D. Mandrioli, "Safety assessment of collaborative robotics through automated formal verification," *IEEE Transactions on Robotics (T-RO)*, vol. 36, no. 1, pp. 42–61, 2019.

[18] O. Birbach, U. Frese, and B. Bäuml, "Rapid calibration of a multi-sensorial humanoid's upper body: An automatic and self-contained approach," *International Journal of Robotics Research (IJRR)*, vol. 34, no. 4-5, pp. 420–436, 2015.

[19] J. H. Lala and R. E. Harper, "Architectural principles for safety-critical real-time applications," *Proceedings of the IEEE*, vol. 82, no. 1, pp. 25–40, 1994.

[20] J. Dobaj, A. Riel, G. Macher, and M. Egretzberger, "Towards devOps for cyber-physical systems (cpss): resilient self-adaptive software for sustainable human-centric smart cps facilitated by digital twins," *Machines*, vol. 11, no. 10, p. 973, 2023.

[21] M. Tacchini, *Functional Safety of Machinery: How to Apply ISO 13849-1 and IEC 62061*. John Wiley & Sons, 2023.

[22] S. Haddadin, S. Haddadin, A. Khoury, T. Rokahr, S. Parusel, R. Burgkart, A. Bicchi, and A. Albu-Schäffer, "On making robots understand safety: Embedding injury knowledge into control," *International Journal of Robotics Research (IJRR)*, vol. 31, no. 13, pp. 1578–1602, 2012.

[23] K. A. Tychola, I. Tsimperidis, and G. A. Papakostas, "On 3D reconstruction using rgb-d cameras," *Digital*, vol. 2, no. 3, pp. 401–421, 2022.

[24] L. Wang, S. Liu, H. Liu, and X. V. Wang, "Overview of Human-Robot Collaboration in Manufacturing," in *Intl. Conf. on the Industry 4.0 Model for Advanced Manufacturing*. Springer, 2020, pp. 15–58.

[25] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[26] "Retenua AB – emitrace® Driver Assistance-Human Workforce Protection with Embedded AI-powered Vision Technology," https://www.retenua.com/en/products/emitrace/, [Online; accessed on 26. Feb. 2025].

[27] R. Mosberger and H. Andreasson, "Estimating the 3d position of humans wearing a reflective vest using a single camera system," in *Field and Service Robotics*. Springer, 2014, pp. 143–157.

[28] R. Mosberger, H. Andreasson, and A. J. Lilienthal, "A customized vision system for tracking humans wearing reflective safety clothing from industrial vehicles and machinery," *Sensors*, vol. 14, no. 10, pp. 17 952–17 980, 2014.

[29] R. Mosberger, B. Leibe, H. Andreasson, and A. J. Lilienthal, "Multi-band Hough Forests for detecting humans with Reflective Safety Clothing from mobile machinery," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 697–703.

[30] C. Schöller, V. Aravantinos, F. Lay, and A. Knoll, "What the constant velocity model can teach us about pedestrian motion prediction," *IEEE Robotics and Automation Letters (RA-L)*, vol. 5, no. 2, pp. 1696–1703, 2020.

[31] S.-L. Sun and Z.-L. Deng, "Multi-sensor optimal information fusion Kalman filter," *Automatica*, vol. 40, no. 6, pp. 1017–1023, 2004.

[32] A. Bourrier, P.-O. Amblard, O. Michel, and C. Jutten, "Multimodal kalman filtering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4413–4417.

[33] N. Mansfeld, M. Hamad, M. Becker, A. G. Marin, and S. Haddadin, "Safety Map: A unified representation for biomechanics impact data and robot instantaneous dynamic properties," *IEEE Robotics and Automation Letters (RA-L)*, vol. 3, no. 3, pp. 1880–1887, 2018.

[34] S. Haddadin, "The Franka Emika Robot: A Standard Platform in Robotics Research," *IEEE Robotics & Automation Magazine (RAM)*, 2024.

[35] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[36] M. Bdiwi, M. Pfeifer, and A. Sterzing, "A new strategy for ensuring human safety during various levels of interaction with industrial robots," *CIRP Annals*, vol. 66, no. 1, pp. 453–456, 2017.