2024 IEEE 20th International Conference on Automation
Science and Engineering (CASE)
August 28 - September 1, 2024. Bari, Italy

Using Multi-channel 3D Lidar for Safe Human-Robot Interaction

Sarthak Arora[1], Kartl

***Abstract*— This paper proposes a novel a channel 3D lidar data for safety in a phy Interaction (pHRI) scenario. To achieve t experiments were conducted to mimic a environment. Data was collected from a participants while performing pre-determin workspace with the robot. A perception pip that leveraged reflectivity images, signal in images, and point-cloud data from a 3D l then used to perform safety based contr the speed and separation monitoring (SSM to support the perception pipeline, a sta detection network was leveraged and fine is provided along with results of the percep based controller.**

I. Introduction

Industry 4.0 has significantly increased the integration of point rich perception sensors into industries including manufacturing, supply chain, warehousing, medical fields, and construction [1]. The integration of these sensors has expanded the automation capabilities of these fields. A key sensor technology integrated across these fields has been lidar. These sensors provide two dimensional and three dimensional information about the environment around them and have been used to detect objects, obstacles, and humans through processes and tasks going on in the workspace around them [2]. With the data provided by lidars, the industry has been able to implement more complex algorithms and autonomous approaches within their fields. This includes the rise of autonomous vehicles, autonomous space vehicle landings, automated guided vehicles (AGVs), unmanned areal vehicles (UAVs), and collaborative robotics applications [3]–[7]. Throughout the past decade, both the algorithms and sensors have continued to see significant innovations.

This material is based upon work supported by the National Science Foundation under Award No. DGE-2125362. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

[1]Sarthak Arora is with the Department of Electrical Engineering and Micro-electronic, Rochester Institute of Technology, 1 Lomb Memorial Drive, Rochester, NY 14623, USA sa9472@rit.edu

[2]Karthik Subramanian is with the Department of Electrical and Micro-electronic Engineering, Rochester Institute of Technology, 1 Lomb Memorial Drive, Rochester, NY 14623, USA kxs8997rit.edu

[3]Odysseus Adamides is with the Department of Electrical and Micro-electronic Engineering, Rochester Institute of Technology, 1 Lomb Memorial Drive, Rochester, NY 14623, USA oaa8092@rit.edu

[4]Ferat Sahin is with Faculty of Electrical and Microelectronic Engineering, Rochester Institute of Technology, 1 Lomb Memorial Drive, Rochester, NY 14623, USA feseee@rit.edu

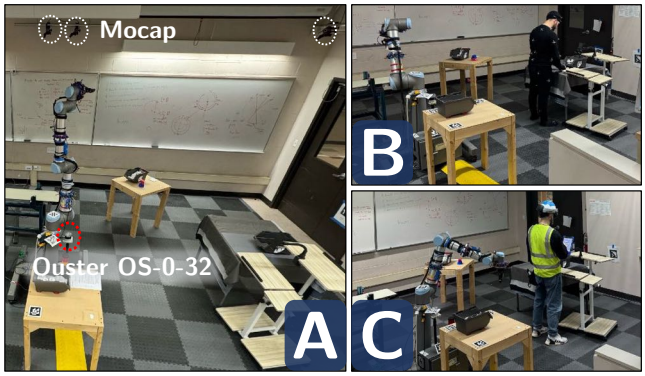


Fig. 1. An image showing different stages of the experiment setup used in this work. In image "A", the layout of the robot workspace is shown along with the exteroceptive sensors used in the setup (encircled in red and white). In image "B", the test subject is wearing a motion capture body suit for acquiring minimum distance associated with the human and robot. In image "C", the participant is wearing a reflective vest and reflective hardhat.

Time-of-Flight (ToF) cameras have begun to increase in depth resolution, it has become easier to calculate depth for stereoscopic cameras, and millimeter wave radar has begun to be used for human tracking applications [8]–[11]. Though there has been a diversification in perception options, lidar remains to be a commonly used sensing method across industrial and research applications [12]–[15]. In the past few years, there has been releases of new lidar products lines which bring new innovations to the perception platform. One product line example was released by the lidar manufacturer Ouster. The "OS" series of lidars includes the OS0, OS1, OS2, and OSDome. Along with various viewing angles and data channels, this product line generates four 2D image data modalities formed from the traditional 3D point cloud the lidar generates [16]. The four frames are range, signal, Near-IR (infra-red), and Reflectivity frames. The Range frame provides a per-pixel ToF distance calculation from the sensor origin to the pixel in the Range frame. The Signal frame provides the light return strength per pixel in the frame. The Near-IR frame provides the light return to the sensor per pixel that was not generated by the laser emitter local to the lidar. This frame measurement is similar to a monochrome IR return from a traditional image sensor. Lastly, the reflectivity frame provides the reflectivity strength per pixel. This frame provides key data on the reflectivity of materials and surfaces in the environment. The significance of Ouster including these frames in addition to the traditional point cloud is that it allows existing 2D machine learning algorithms to be directly applied to the 3D lidar data [16]. Hence, by leveraging the data provided by the channels, we aim to make the following contributions:

1) Develop a lidar based dataset with multiple participants with varying clothing and body shapes in realistic shop-floor conditions.
2) Demonstrate a successful use of the data collected by training a state-of-the-art object detector with validation and testing.
3) Propose an improved formulation of the algorithm presented in [17] to compute the directed robot velocity.

II. Literature Survey

2D frames of 3D lidar point clouds have been used in a number of research fields. This includes [18], where the reflectivity image was used to correct drone odometry. Additionally, [19] fuses the multiple modalities to increase 3D object detection performance. These alternate frame data formats have also been used in the automotive field to test segmentation of humans, vehicles, and other traffic objects without the use of a traditional CMOS image sensors [20]. In these different applications, there are plenty of previous works that illustrate a sufficient approach for feature extraction and data formatting to feed image based classifiers and algorithms. With the dawn of Industry 5.0, it is imperative for lidar to maintain compatibility with 2D machine vision and machine learning algorithms such that lidars match the performance of other perception systems used across industry [21]. Industry 4.0 setup the infrastructure of digitally driven and automated processes, Industry 5.0 pushes researchers to look deeper at these processes and their impact on the human individuals who must coexist with this infrastructure. A key research area that will continue to be a focus area in Industry 5.0 is Human Robot Collaboration (HRC). In this field, the pose of the worker, distance from worker to robot, and trajectories of the human and robot in the workspace are vital to increasing the safety and comfort of the worker [17], [22], [23]. Speed and separation monitoring (SSM) is one of the four major collaborative approaches identified in the International Organization for Standardization (ISO) standard ISO/TS 15066:2016 [24]. In the field of SSM research, a number of different sensor configurations and modalities are considered including ToF cameras, stereo cameras, mmWave radars, ultrasonic sensors, and lidars [23]. Lidar was the primary sensor used in the early years of SSM research [25]. As innovations in computation and perception have progressed, the other perception modalities have seen a rise of use in the field. To track the human in the scene, it is crucial for the image based perception systems used in an SSM setup to feed data to convolutional neural networks (CNNs) [26]. This localization of the human in the frame enables the computation of minimum distance data needed for an SSM algorithm. With the Ouster OS-0-32, the lidar data can also be used to directly feed CNN based algorithms for human position, and pose tracking.

In this paper, frame based lidar data is directly used to train a YOLOv9 [27] model in contrast to traditional methods which require raw 3D point cloud processing and mapping prior to the input into a neural network. Additionally, the data captured in this work consists of diverse body shapes and clothing material in an industrial environment. Furthermore, the data and model is applied to a simple, generalized SSM algorithm which outputs a safety distance and an operational velocity scaling factor. Lastly, the paper explores the viability of a vertical and horizontal field of view (FoV) lidars for safety based applications. The dataset, and trained model will be shared in the future for the research community.

III. Methodology

This section covers the various components involved in the experimental process illustrated in Fig. 1. The goal of of the setup was to explore the usage scenarios for 3D lidars such as the Ouster OS-0-32 in an industrial shop floor environment. This environment was comprised of mostly static objects (workbenches etc.) with a limited number of dynamic objects (humans & robots) within the lidar FoV.

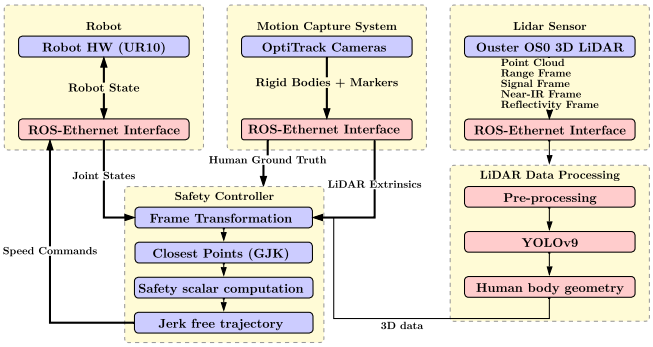


Fig. 2. Control schema showing the complete system, our communication is powered by Robot Operating System (ROS).

A. Setup & Calibration

In this step, the focus was on achieving a time synchronization between the heterogeneous data streams emanating from the lidar sensor, motion capture, and the robot control box as shown in Fig. 2. Synchronization relied on a local high speed Ethernet based network that exhibited an average latency of approximately 0.25 milliseconds (round-trip time). Therefore, it was assumed that the delta time between the time of arrival of data packets and the time of origin was negligible. Finally, an asynchronous time synchronization on the various streams was performed. The inter-stream time delta of the synchronization was 5 milliseconds. In the calibration procedure, the main goal was to obtain the rigid body transformations of all the sensing data in a common reference frame. Customized rigid body marker-sets we developed for the motion capture system. These marker were affixed to the lidar, pedestal of the robot, and also on the skeleton tracking body suit worn by a human participant. This step provided a coarse calibration, however, for a better estimate of the lidar extrinsics an optimization based point-set alignment was used as shown in [28].

B. Data Collection and Labeling

Once a synchronized and calibrated setup was achieved, the incoming data was recorded over the network on a local disk storage. The following data fields were focus on:

- **Lidar Data** at 20 Hz: Point-cloud, Refelectivity, Signal, Near-Infrared and Stacked images
- **Motion Capture Data** at 120 Hz: Rigid bodies and 3D

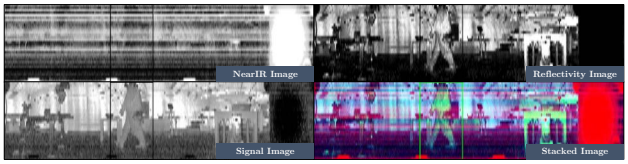


Fig. 3. Clean samples of the reflectivity, signal, near-IR, and a depth-wise stacked image of the first three. The annotation is overlayed on the grayscale images in black and in green on colored images.

After the data was collected, it was then processed for labeling and downstream tasks such as low-level image processing and classification as seen in Fig. 3. The bulk of processing comprised of pre-processing steps applied to the lidar point-cloud, and image quadruplet obtained from the lidar data. First, the lidar point-cloud images were "*destaggered*" as mentioned in the documentation provided by the manufacturer in [29]. The main idea behind this step was to remove the time offset from each element of the lidar data (point-cloud and images). Afterwards, the image quadruplet was subjected to bit depth down-sampling from 16-bit to 8-bit image data. As the image resolution was 1024×32, the images had to be resized to 1024×256 by applying bi-linear interpolation. The images were then subjected to auto-exposure adjustment and histogram equalization as part of pre-processing. The images were then annotated in a semi-automated fashion with bounding boxes in MSCOCO format [30]. For semi-automation, the static nature of the environment was exploited to remove a large number of points by background removal and applying statistical outlier rejection on the remaining points. Afterwards, noisy bounding-box labels were generated by re-projecting the non-stationary 3D points into the images. Ultimately, the bounding boxes were hand tuned.

C. Network Training and Inference

The YOLOv9 [27] object detection network was selected to annotate bounding boxes around the human body shape in the lidar data (to image quadruplet only). For this step, two datasets were developed from the data collected during multiple experiments. The two variants comprised of single-channel annotated reflectivity images and multi-channel annotated images where reflectivity, signal, and near-infrared images were stacked depth-wise forming a tensor. It must be noted, that the images in the datasets represented only a subset of the total data recorded during the experiments. A larger full version of the dataset will be made available for the research community. Fine tuning was performed on a pre-trained variant of YOLOv9 called "*YOLOv9-C*" that had fewer parameters than the largest YOLOv9 variant called "*YOLOv9-E*". The network was selected due to it's state-of-the-art performance and efficiency as shown in [27]. Both datasets consisted of 14,000 annotated images, the dataset split was selected as 80% training and 20% validation. For testing, new dual variants of single-channel and multi-channel datasets (comprised of unseen data by the network) were prepared. As safety is one of the key challenges in pHRI [22], it is important to analyze every labeled and unlabeled image by the network at inference time to determine its suitability in a high stakes scenarios. Therefore, the test-set created was representative of one full trial performed by a human subject during the experiment, and validated using the fine-tuned network. For training, the stochastic gradient descent (SGD) optimizer with a momentum of 0.937 was selected. The batch size and epochs were chosen to be 16 and 50, respectively.

D. Human point cloud extraction

In this phase of the pipeline, the annotations provided by the aforementioned network at inference time were used. As the spatial structure of the lidar frame (comprising of point-cloud and image quadruplets) allowed for a bi-directional mapping between the images and the point-cloud. The bounding-box rectangles were projected into a corresponding point-cloud and points external to the region of interest were pruned as seen in Fig. 4. This reduced the total number of points from 1024×32 to approximately 20×50 (based on the largest possible size of the bounding box). Then, plane-segmentation and DBSCAN [31] clustering were used to extract the points associated with the human body shape.

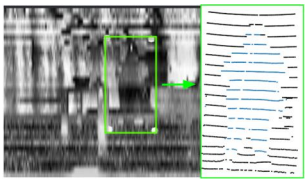


Fig. 4. An image showing human shape geometry extraction using a point-cloud along with a bounding determined from a reflectivity image. Points in blue represent the human body, points in black are rejected as background.

E. Speed and Separation Monitoring Algorithm

The method towards the implementation of the Speed and Separation Monitoring (SSM) was derived from [17]. This work defines the 3D geometrical (in a common reference frame as the robot) representation of the human operator in the workspace. A scene graph was constructed with convex decomposed geometries and closest pair of point queries between the human and the robot were performed. The algorithm of choice for such tasks was the "GJK" algorithm [32] which is widely used for such applications. The closest pair of points allowed for the computation of the minimum distance vector. This vector was used to compute the protective safety distance (threshold or a barrier around the robot) to trigger the robot stop behavior. As stated in [24] and [17], the speed and separation monitoring equation is given by (1).

$$S_{safety}(t_0) = V_{human}\cdot(t_r+t_s)+V_{robot}\cdot t_r+C+Z_s+Z_r \quad (1)$$

Z_s & Z_r represent the position measurement uncertainties for human and robot respectively. These values were obtained from the datasheets of the robot and the exteroceptive sensing equipment used. C is the intrusion distance which is defined by [33]. In essence, it represents the threshold at which an obstacle is successfully detected. t_r and t_s represent the control loop processing time and the time required by the robot to come to a full stop, respectively. These time values could also be obtained from the robot's (UR10) datasheet or empirically estimated. The robot stopping time t_s, can also be tuned but should be lower bound by the worst case stopping time. To achieve jerk free stop behavior, the online trajectory generation library [34] was used.

As governed by the standard [24], most terms in the equation can be substituted with constants, the only quantity which is non-trivial to compute is V_{robot}. Thus, Algorithm 1 is proposed to compute the directed velocity of the robot towards the human based on velocity kinematics of the robot:

Algorithm 1: Algorithm to compute directed $\mathbf{V}_{robot}$

Input: P_{human}, P_{robot}, W_{max}
Routines Used: $jacobian$,is_valid,$normalize$
Output: V_{robot} in direction of the operator
$\vec{S} \leftarrow P_{human} - P_{robot}$
$\vec{z} = (0,0,1)^T$
if *is_valid*($\vec{S}$) **then**
  $\vec{f} = normalize(\vec{S})$
  $\vec{r} = \vec{z} \times \vec{f}$
  $\vec{u} = \vec{f} \times \vec{r}$
  ${}^{w}\mathbf{T}_r = \begin{pmatrix} \vec{\mathbf{r}} & \vec{\mathbf{u}} & \vec{\mathbf{f}} & \mathbf{P}_{robot} \\ 0 & 0 & 0 & 1 \end{pmatrix}$
  $\mathbf{J}_{robot} = jacobian(\mathbf{q}, {}^{w}\mathbf{T}_r)$
  $\left[{}^{w}\mathbf{v}_{twist}\right]_{6\times1} = \mathbf{J}_{robot} \cdot \mathbf{q}$
  ${}^{w}\mathbf{T}_h = {}^{w}\mathbf{T}_r \cdot \begin{pmatrix} 1&0&0&0 \\ 0&1&0&0 \\ 0&0&1&\|\vec{\mathbf{S}}\| \\ 0&0&0&1 \end{pmatrix}$
  ${}^{h}\mathbf{v}_r = {}^{w}\mathbf{T}_h{}^{-1} \cdot \left[{}^{w}\mathbf{v}_{twist}\right]_{4\times4}$
  $V_{robot} = sign({}^{h}\mathbf{v}_r\langle\mathbf{z}\rangle) \cdot \|{}^{h}\mathbf{v}_r\langle\mathbf{z}\rangle\|$
  $return\ V_{robot}$
else
  $return\ NaN$

$\vec{S}$ is the the human-robot minimum distance vector computed from the 3D point pair P_{human} & P_{robot}. $\vec{f}$, $\vec{r}$ & $\vec{u}$ are normalized vectors representing the right-handed coordinate system. ${}^{w}\mathbf{T}_r$ is a homogeneous transform representing P_{robot} along the z-axis (pointing forward) in world (w) coordinates. ${}^{w}\mathbf{v}_{twist}$ is the instantaneous twist directly computed from the joint velocities $\mathbf{q}$ of the robot and jacobian $\mathbf{J}_{robot}$ from joint positions $\mathbf{q}$. It is then represented in matrix form to compute the twist frame in the direction of the human. Ultimately, V_{robot} is computed by using the elements along the z-axis as a signed scalar. Finally, V_{robot} can be substituted in (1) to obtain the safety distance, S_{safety}. The magnitude of minimum distance vector $\|\vec{S}\|$ and S_{safety} can be used to compute the speed scaling commands sent to the robot controller by (2). As compared to [17], Algorithm 1 presents a more general formulation that computes the instantaneous directed velocity using the joint velocities instead of leveraging link velocities. This approach also bypasses ellipsoid approximations of the robot geometry as done in [17]. Furthermore, the human-robot minimum distance is directly computed with respect to the entire shape geometry of the robot. While in [17], the minimum distance was computed between the human and the link centroids of the robot. Furthermore, for computing the term to scale the robot's operational speed the following formulation was used

$$\rho_{scaling} = \frac{max(\|\vec{S}\| - S_{safety}, 0)}{W_{max}} \quad (2)$$

Such that $\rho_{scaling} \in [0,1]$ and W_{max} is the distance limit in meters for the robot workspace beyond which sensor readings are clamped. $\rho_{scaling}$ can be used to uniformly scale the joint velocities $\mathbf{q}$ of the robot also shown in [17].

IV. Experiment

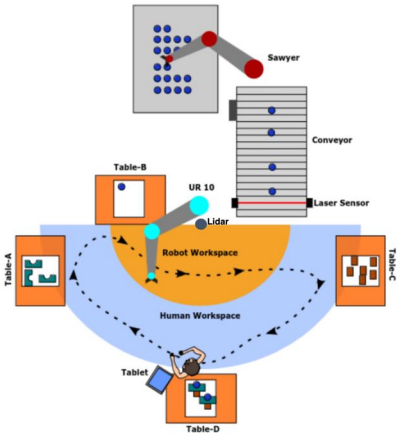


Fig. 5. Flow diagram of the experiment performed as prescribed by the authors in [35]

Fig. 5 depicts the experimental configuration of distinct tasks allocated to both the human and the robot in an assembly line. Specifically, a collaborative robot was responsible for extracting a component from a pallet and depositing it onto a conveyor belt. Subsequently, a UR-10 robot retrieved the necessary component from the conveyor and situated it within a bin accessible to the human operator, both the human and UR-10 coexisted within the same workspace. There were four stations in the shared workspace. The human worker walked to each of these stations completing the assembly of a PVC coupling. One of these parts was provided by UR-10. Once the assembly of a single part was completed, the human was required to deposit the item at a designated location and perform an arbitrary task at the same station. This process was repeated 24 times in one trial. A lidar captured point clouds and multi-channel imagery throughout the experiment. Each worker participated in 6

trials, in 3 of those trials, they were required to wear a high visibility jacket. Additionally, the entire workspace was monitored with a motion capture system which possessed 13 cameras that flooded the workspace with 940 nm infrared light.

V. Results and Discussion

A. Dataset

The experiment involved 17 participants, 29% of the participants were of female sex and the remaining 71% were of male sex. The variety of clothing, fabrics, and colors worn by participants were recorded. A world-cloud image to represent this diversity is illustrated in Fig. 6. The most common clothing color, fabric, and type were black, cotton and jeans with hoodie, respectively. This gave a minor insight that clothing worn by operators in shop-floors could also be dark in color and of cotton material. Hence, a lidar should be able to detect these materials based on any arbitrary surface reflectivity exhibited by them.

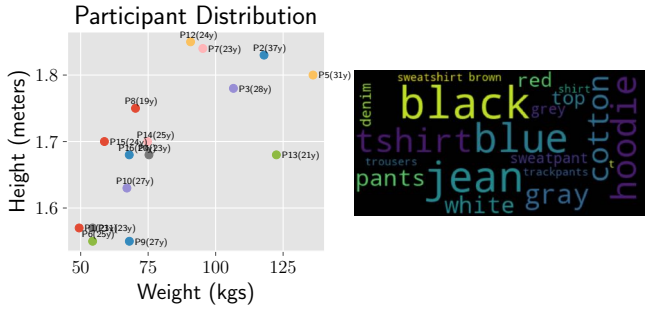


Fig. 6. A plot and word-cloud showing the participant attribute distribution. Left: the weight-height and age distribution of the participants. Right: a word cloud showing a distribution of clothing types, materials and colors.

The weight of the participants ranged from 49 kg to 136 kg, and the height ranged from 1.5 m to 1.85 m as shown in Fig. 6. It is vital for any learning based model to be aware of varying body geometries in a shop-floor environment. The approval for Human subject research was granted by Rochester Institute of Technology. (Approval number: **21081267**)

After the data collection, pre-processing was applied and the training and validation sets (for "single-channel" and "multi-channel") were prepared with about 12,000 and 2200 images, respectively. Fig. 7 is an example of images fetched from the dataset.

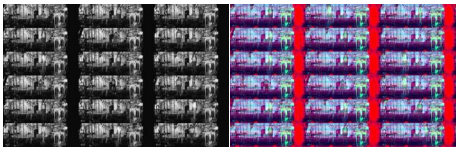


Fig. 7. Tiled layout of 18 samples randomly drawn from each training dataset. Left: a snapshot of the single-channeled dataset built with reflectivity images. Right: a snapshot of the multi-channeled dataset built with depth-wise concatenation of reflectivity, signal, and near-infrared images.

B. Quantitative Results

The two previously mentioned datasets were used to train the YOLOv9 object detector in a binary detection mode. The validation curves during training session are shown in Fig. 8. During the "multi-channel" training it was observed that the YOLOv9 network converged faster and exhibited a higher mAP50-95 validation score.

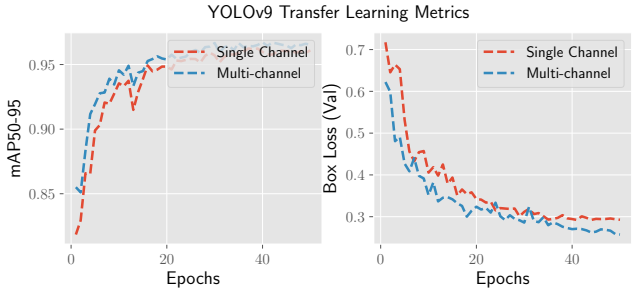
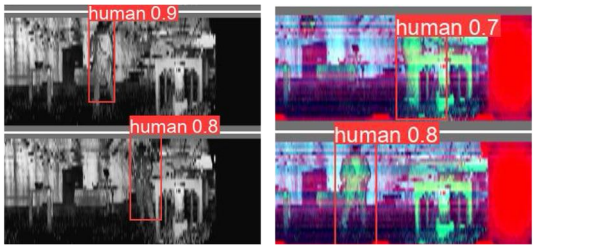


Fig. 8. Plots showing the validation metrics during YOLOv9 fine-tuning on the datasets prepared namely Single-channel & Multi-channel datasets.

After training the network, inference was performed on unseen lidar sequences of 12,500 samples for both variants. In Fig. 9, it was observed that the "multi-channel" variant performed approximately 1% better than the "single-channel" variant. However, it was noted that the classifier confidence during inference was more robust during "single-channel" inference. On analyzing the spread of the confidence values of the classifiers, it was found that the multi-channel detector was measurably less certain than the single-channel variant.



		Single Channel		Multi-channel	
		Human	Background	Human	Background
Predicted	Human	**11549 (98%)**	184	**11685 (99%)**	392
	Background	207 (2%)	NA	67 (1.0%)	NA
		Actual			

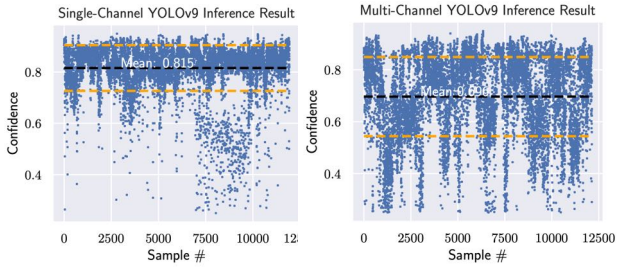


Fig. 9. Figures showing inference examples, confusion matrix, and confidence scatter plots on the test sets.

To measure the accuracy of the lidar for pHRI scenarios, the closest pair of points between the human operator and the robot were recorded with the lidar and the motion capture

system as seen in Fig. 10. For the on-robot base mounted 3D lidar, the root mean square error (RMSE) was more than 4 times lower than on-robot time-of-flight sensing rings in [17]. The margin of error was found to be lower bounded by 3mm as reported by the manufacturer.

	Lidar	ToF Rings [17]
RMSE (m)	**0.0605**	0.25

TABLE I

RMSE COMPARISON FOR LIDAR AND TIME-OF-FLIGHT SENSING RINGS

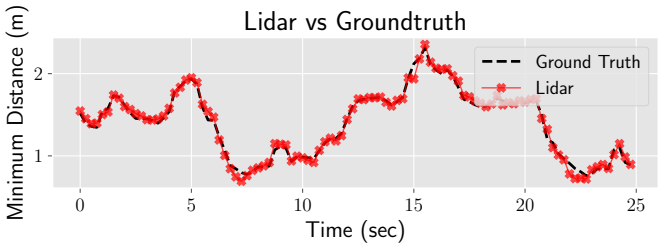


Fig. 10. Plot showing the minimum-distance comparison between data acquired from the lidar and motion capture system overlay-ed on top of each other.

The results from the safety algorithm are presented in Fig. 11. A 25 $second(s)$ long recording of the results was analyzed; the directed robot velocity V_{robot} was found to be proportionally tracked by safety distance S_{safety} due to its linear dependence on the prior. V_{human} was set to $1.6m/s$ (prescribed by [24]) and the remaining terms in (1) were construed from the robot's datasheet. Between 18.5 and 19.0 seconds, the speed scaling term $\rho_{scaling}$ decayed when the minimum distance $\|\vec{S}\|$ violated or tended towards S_{safety}. It should also be noted that $\rho_{scaling}$ was always below 0.5, as the human subject was always within 1.5 meters ($< W_{max}$) of the robot. Furthermore, even though there was no smoothing and filtering applied to the data, $\|\vec{S}\|$ computed from lidar data was smoother than in [17], where an exponential filter was used. As filtering can introduce time delay in the controller, which is considered a risk in safety-centric scenarios. Therefore, leveraging a system (lidar) that can provide data at high sampling frequencies and low noise is vital.

C. Qualitative Results

During data collection, it was observed that the response of the lidar was poorer in certain scenarios where the participants were wearing significantly darker clothes even in close proximity to the robot. It was found that the reflectivity and range images provided by the lidar exhibited the presence of holes. As a consequence, the point cloud lacked the 3D information associated with the human's shape geometry (points were missing from the point cloud). This phenomenon is illustrated in a side by side comparison shown in Fig. 12. It should be noted that in the left half of the figure, the participant was wearing a high-reflectivity vest with black cotton garments underneath. Only the points associated with the reflective vest were reported by the lidar.

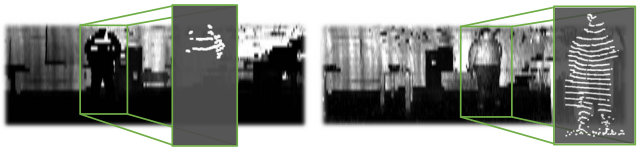


Fig. 12. On left, lidar reflectivity image with holes with its corresponding 3D point-cloud in perspective view. On right, a healthy sample of the reflectivity image with its 3D point cloud.

Another limitation was observed, wherein the "multi-channel" variant performed poorly after the floor layout changed. This limitation is shown in Fig. 13, this can be explained due to a distribution shift, the network is biased directly on the metrics associated with the photons scattered in the environment. The colored patches in the image can also create ambiguous textures that can confuse the network. It should be noted, that in this image the participant is wearing a reflective vest. This can be also be addressed by a higher resolution lidar such as OS-0-128 where the base resolution is 1024×128, hence the effect of up-scaling will not create aliasing artifacts.

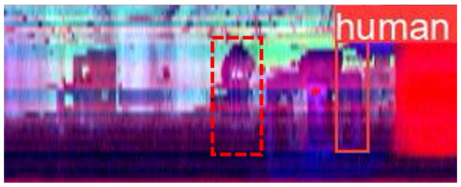


Fig. 13. A miss-classification performed by the "multi-channel" variant due to room layout change and ambiguous texture patches in the input image. The correct bounding box is drawn in a dashed bounding box.

Fig. 14 shows the monotonic nature of the lidar recordings on the left. This can cause a rolling shutter effect; while the lidar records at relatively high frequency (at 20 Hz), it may be susceptible to creating artifacts if an object in the frame moves fast enough from one point to another before the entire frame is scanned. This may create a situation where the moving object appears to have teleported in the recorded frame. Swiftly moving objects can also appear distorted as they may have been recorded at staggered intervals.

To hypothesize an explanation for better validation and inference performance by the "multi-channel" variant, the structured similarity index metric (SSIM) matrix was used. As shown on the right hand side in Fig. 14, the relative SSIM of each image type is significantly below 1.0, there is a likelihood that the network can extract additional features from the added channels. If the features were redundant, the relative SSIM (the matrix elements would diffuse more) would be closer to 1.0. On the other hand, the Near-IR channel nominally tends to exhibit higher ambient noise than other images. This could explain a much lower SSIM value for Near-IR with respect to reflectivity and signal channels.

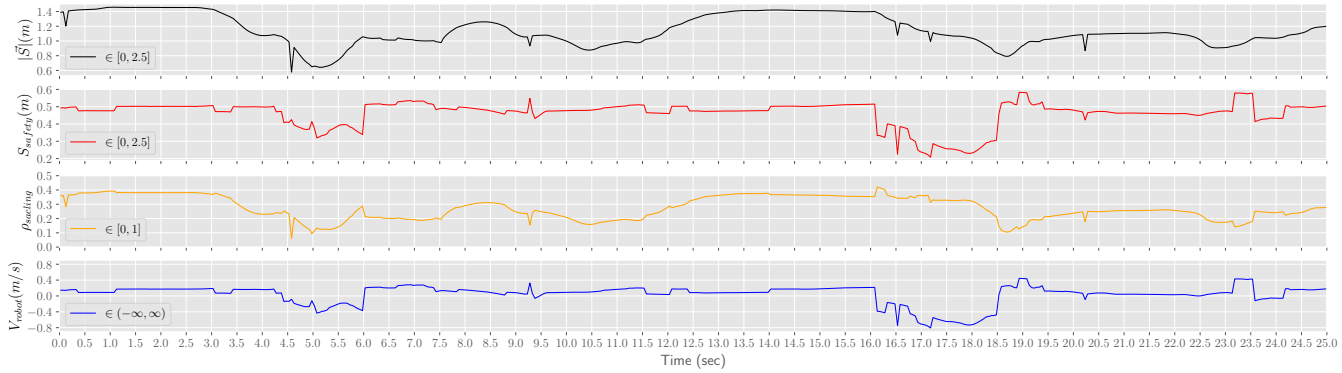


Fig. 11. Time series (25 seconds) plots showing the results of the directed robot velocity computation (bottom most in blue) towards the human along with the minimum distance (top in black), safety distance threshold (second from top in red) computed with the SSM Equation [24] and the speed scaling factor (third from top in yellow) used for modulating the operational speed of the robot.

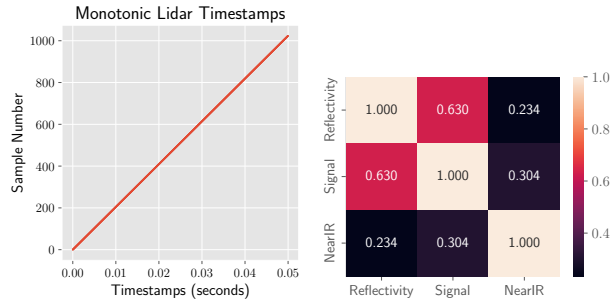


Fig. 14. Left, the timestamps of the sequential readout perform by the lidar. Right, a matrix showing the relative structured similarity index metric (SSIM) of the lidar images.

VI. Conclusions & Future Work

The on-robot base mounted lidar can significantly outperform on-robot time-of-flight sensing rings due to the 3D point-cloud and 2D image data. Furthermore, the bi-directional $2D \Leftrightarrow 3D$ mapping enables for higher level tasks such as object detection on images and subsequently, region-of-interest extraction on corresponding point-clouds. This leads to a more efficient perception pipeline as image based backend(s) can be used to bootstrap detection networks while pruning the 3D search space. Also, due to this capability, we were able to semi-automate bounding box annotation for our datasets. In the future, this can enable the application of techniques such as continual learning [36].

The lidar also exhibits some limitations due to the presence of holes in the image channels which affect the quality of the point-cloud. Therefore, in a shop-floor it is vital to wear high-reflective markers such as vest and helmets as they can alleviate the presence of holes in the lidar data. Ultimately, we can conclude that the use of the 3D lidar in close proximity pHRI scenarios is viable, as long as steps are taken to prevent sensing failures and pitfalls.

For future works, the first step is be to develop a target network that can directly handle the input sizes provided by the lidar and is designed to work with 16-bit precision. To overcome the distribution shift problem, the channel order can be randomized while also introducing small changes in the floor layout so that the network becomes more robust. The changes required would be quite small, as a shop-floor environment is more static than an outdoor scenario. Exploring deep learning based image up-scaling techniques such as [37] and usage of more advanced sensing hardware (OS-0-128) will also provide us with more reliable inference. Another downstream task that we are already working is instance segmentation, we are currently working on developing mask annotation for the lidar images.

For the safety controller, leveraging directed velocity of the human operator towards the robot will also aid the safety barrier to be relaxed in situations where the human is moving away from the robot. As we assume V_{human} to be a positive constant, it implies that operator is always moving in the direction of the robot with a constant velocity. Therefore, measuring the velocity of the operator in real-time will be beneficial for robot productivity without sacrificing operator safety.

References

[1] Li Da Xu, Eric L. Xu, and Ling Li. Industry 4.0: state of the art and future trends: International Journal of Production Research. *International Journal of Production Research*, 56(8):2941–2962, April 2018. Publisher: Taylor & Francis Ltd.

[2] Paul F. McManamon and Society of Photo-optical Instrumentation Engineers. *LiDAR technologies and systems*, volume PM300. SPIE, Bellingham, Washington (1000 20th St. Bellingham WA 98225-6705 USA), 2019.

[3] G. Ajay Kumar, Ashok Kumar Patil, Rekha Patil, Seong Sill Park, and Young Ho Chai. A LiDAR and IMU Integrated Indoor Navigation System for UAVs and Its Application in Real-Time Pipeline Classification. *Sensors (Basel, Switzerland)*, 17(6):1268, June 2017.

[4] Heng Wang, Bin Wang, Bingbing Liu, Xiaoli Meng, and Guanghong Yang. Pedestrian recognition and tracking using 3D LiDAR for autonomous vehicle. *Robotics and Autonomous Systems*, 88:71–78, February 2017.

[5] Charith Munasinghe, Fatemeh Mohammadi Amin, Davide Scaramuzza, and Hans Wernher van de Venn. COVERED, CollabOratiVE Robot Environment Dataset for 3D Semantic segmentation. In *2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA)*, pages 1–4, September 2022.

[6] Zoltan Rozsa and Tamas Sziranyi. Obstacle Prediction for Automated Guided Vehicles Based on Point Clouds Measured by a Tilted LIDAR Sensor. *IEEE Transactions on Intelligent Transportation Systems*, 19(8):2708–2720, August 2018. Conference Name: IEEE Transactions on Intelligent Transportation Systems.

[7] Răzvan-Ionuț Bălașa, Ghoerghe Olaru, Daniel Constantin, Amado Ștefan, Ciprian-Marian Bîlu, and Maria Beatrice Bălăceanu. LIDAR based distance estimation for emergency use terrestrial autonomous robot. In *2021 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–4, July 2021.

[8] Jeffrey Too Chuan Tan and Tamio Arai. Triple stereo vision system for safety monitoring of human-robot collaboration in cellular manufacturing. In *2011 IEEE International Symposium on Assembly and Manufacturing (ISAM)*, pages 1–6, May 2011.

[9] Odysseus Alexander Adamides, Alexander Avery, Karthik Subramanian, and Ferat Sahin. Evaluation of On-Robot Depth Sensors for Industrial Robotics. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1014–1021, October 2023. ISSN: 2577-1655.

[10] Bakir Lacevic, Andrea Maria Zanchettin, and Paolo Rocco. Safe Human-Robot Collaboration via Collision Checking and Explicit Representation of Danger Zones. *IEEE Transactions on Automation Science and Engineering*, 20(2):846–861, April 2023. Conference Name: IEEE Transactions on Automation Science and Engineering.

[11] Barnaba Ubezio, Christian Schöffmann, Lucas Wohlhart, Stephan Mülbacher-Karrer, Hubert Zangl, and Michael Hofbaur. Radar Based Target Tracking and Classification for Efficient Robot Speed Control in Fenceless Environments. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 799–806, September 2021. ISSN: 2153-0866.

[12] Arne Peters, Adam Schmidt, and Alois C. Knoll. Extrinsic Calibration of an Eye-In-Hand 2D LiDAR Sensor in Unstructured Environments Using ICP. *IEEE Robotics and Automation Letters*, 5(2):929–936, April 2020. Conference Name: IEEE Robotics and Automation Letters.

[13] Aquib Rashid, Kannan Peesapati, Mohamad Bdiwi, Sebastian Krusche, Wolfram Hardt, and Matthias Putz. Local and Global Sensors for Collision Avoidance. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 354–359, September 2020.

[14] Himansu Sekhar Behera and Anandakumar M Ramiya. Urban flood modelling simulation with 3D building models from airborne LiDAR point cloud. In *2022 IEEE Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS)*, pages 145–148, March 2022.

[15] Efstathios Karypidis, Georgios Zamanakos, Lazaros Tsochatzidis, and Ioannis Pratikakis. Point Contrastive learning for LiDAR-based 3D object detection in autonomous driving. In *2023 24th International Conference on Digital Signal Processing (DSP)*, pages 1–5, June 2023. ISSN: 2165-3577.

[16] Fisher Shi. Object Detection and Tracking using Deep Learning and Ouster Python SDK | Ouster, March 2022.

[17] Shitij Kumar, Sarthak Arora, and Ferat Sahin. Speed and Separation Monitoring using On-Robot Time-of-Flight Laser-ranging Sensor Arrays. In *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, pages 1684–1691, August 2019. ISSN: 2161-8089.

[18] Yanfeng Zhang, Yunong Tian, Wanguo Wang, Guodong Yang, Zhishuo Li, Fengshui Jing, and Min Tan. RI-LIO: Reflectivity Image Assisted Tightly-Coupled LiDAR-Inertial Odometry. *IEEE Robotics and Automation Letters*, 8(3):1802–1809, March 2023. Conference Name: IEEE Robotics and Automation Letters.

[19] Yunze Man, Xinshuo Weng, Prasanna Kumar Sivakumar, Matthew O'Toole, and Kris Kitani. Multi-Echo LiDAR for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3743–3752, October 2021. ISSN: 2380-7504.

[20] Maria Tsiourva and Christos Papachristos. LiDAR Imaging-based Attentive Perception. In *2020 International Conference on Unmanned Aircraft Systems (ICUAS)*, pages 622–626, September 2020. ISSN: 2575-7296.

[21] João Barata and Ina Kayser. Industry 5.0 – Past, Present, and Near Future. *Procedia Computer Science*, 219:778–788, January 2023.

[22] Valeria Villani, Fabio Pini, Francesco Leali, and Cristian Secchi. Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications. *Mechatronics*, 55:248–266, November 2018.

[23] Shitij Kumar, Celal Savur, and Ferat Sahin. Survey of Human–Robot Collaboration in Industrial Settings: Awareness, Intelligence, and Compliance. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(1):280–297, January 2021. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics: Systems.

[24] ISO. ISO/TS 15066:2016(en) Robots and robotic devices — Collaborative robots, 2022.

[25] Jeremy A. Marvel and Rick Norcross. Implementing speed and separation monitoring in collaborative robot workcells. *Robotics and computer-integrated manufacturing*, 44(Journal Article):144–155, 2017. Place: OXFORD Publisher: Elsevier Ltd.

[26] Justyna Patalas-Maliszewska, Adam Dudek, Grzegorz Pajak, and Iwona Pajak. Working toward Solving Safety Issues in Human–Robot Collaboration: A Case Study for Recognising Collisions Using Machine Learning Algorithms. *Electronics (Basel)*, 13(4):731, 2024. Place: Basel Publisher: MDPI AG.

[27] Chien-Yao Wang, I.-Hau Yeh, and Hong-Yuan Mark Liao. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information, February 2024. arXiv:2402.13616 [cs].

[28] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, April 1991. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[29] Ouster. Ouster SDK — Ouster Sensor SDK 0.10.0 documentation, 2022.

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context, February 2015. arXiv:1405.0312 [cs].

[31] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231, Portland, Oregon, August 1996. AAAI Press.

[32] E.G. Gilbert, D.W. Johnson, and S.S. Keerthi. A fast procedure for computing the distance between complex objects in three-dimensional space. *IEEE Journal on Robotics and Automation*, 4(2):193–203, April 1988. Conference Name: IEEE Journal on Robotics and Automation.

[33] ISO. ISO 13855:2010(en) Safety of machinery — Positioning of safeguards with respect to the approach speeds of parts of the human body, 2010.

[34] Lars Berscheid and Torsten Kröger. Jerk-limited Real-time Trajectory Generation with Arbitrary Target States, June 2021. arXiv:2105.04830 [cs].

[35] Celal Savur. *A physiological computing system to improve human-robot collaboration by using human comfort index*. PhD thesis, Rochester Institute of Technology, Rochester, NY, 2022. Dissertation/Thesis.

[36] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A Comprehensive Survey of Continual Learning: Theory, Method and Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2024. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[37] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced Deep Residual Networks for Single Image Super-Resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140, July 2017. ISSN: 2160-7516.