Multi-3DSA: A Multi-resolution 3D Mapping for Safety in Automated Factories

Lucas Carvalho de Lima[1,2], Sisi Liang[2], Paul Flick[2], Michael Brünig[1], Nicholas Panitz[2], Jen Jen Chung[1,2]
[1]The University of Queensland, [2]CSIRO Technology, Australia
Email: l.carvalhodelima@uq.edu.au

***Abstract*—Ensuring the safety of humans in mixed autonomy factories where the operation area is shared amongst workers, mobile robots, and forklifts remains a challenge. Current systems rely on inherently myopic sensing onboard mobile agents or human supervision for accident prevention, which are costly and prone to errors due to blind corners. We address these challenges with a novel system, Multi-3DSA, that leverages the surveillance camera network of factories to reconstruct and track agents in the scene, further detecting accident risks and providing safety actions accordingly. Specifically, we perform a 3D reconstruction from vision-only that uses multi-view shape carving integrated with multi-resolution mapping based on occupancy and wavelet compression. The multi-scale 3D maps are then used by a supervisory system that tracks classified objects (humans, AGVs, forklifts) and finally sends safety triggers after identifying collision risks. The system was evaluated on recorded data and deployed on a real test bench, showing the effectiveness of the 3D reconstruction and monitoring capabilities.**

***Index Terms*—3D Reconstruction; Robotics and Automation in Manufacturing; Mapping**

I. Introduction

Human-robot collaboration is key to achieving the productivity benefits of Industry 4.0. In these smart factories, the seamless integration of autonomous robots and workers requires complex coordination and robust safety systems to prevent accidents. Although AGVs (Autonomous Guided Vehicles) and forklifts are commonly equipped with LiDAR [1], radar, or ultrasonic sensors for obstacle detection, factory environments remain challenging due to cluttered layouts, blind corners, and large loads that can obstruct sensors. Alternatively, safety is also ensured by dedicated monitoring personnel. These solutions are prone to accidents due to low visibility, incur higher costs, and have poor scalability.

To address the challenges of improving safety in mixed autonomy factories, we present Multi-3DSA, a 3D situational awareness solution that performs real-time monitoring and tracking of humans, AGVs and forklifts, identifying potential collisions and providing alerts for reliable human safety on the shared factory floor. Our solution uses the existing network of surveillance cameras installed a priori in most factories as the main sensing source, effectively reducing installation and maintenance costs. Our system first performs 3D reconstruction of the factory scene from images using a shape carving technique and further extracts semantic information, allowing for object identification and tracking. Then, we use online location estimates and semantic data to provide safety warnings to humans and automated AGVs.

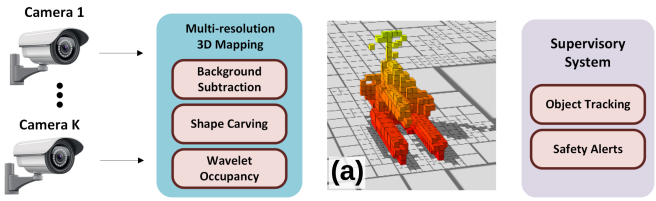


Fig. 1: Overview of our proposed multi-resolution mapping and safety system. After background/foreground separation, a shape-carving method integrated with wavelet occupancy is performed to reconstruct the scene. The supervisory system then employs semantic classification and tracking of objects to send safety alerts and stops to prevent accidents. (a) Shows in 3D a reconstructed pallet jack.

The reconstructed environment is represented with a grid-based voxel map; however, instead of using fixed-sized voxels, whose computational requirements scale linearly with environment size, we introduce a novel mapping method that combines shape carving reconstruction and multi-resolution wavelet decomposition. The multi-resolution map automatically adapts to the necessary level of detail in the scene (see Fig. 1) and performs fast computations, adhering to the strict runtime requirements of the safety task.

The main contributions of this work are the following:

- Multi-3DSA, a comprehensive method for real-time situational awareness and risk management in smart factories;
- A shape-carving 3D reconstruction method with a multi-resolution mapping representation based on wavelet decomposition for higher scalability; and
- Evaluations with real data including deployment in a factory-like environment.

II. Related Work

A. Human-Robot Safety in Factories

Traditional methods to prevent accidents between mobile agents (AGV and forklifts) and humans rely on physical separation [2] and onboard sensing (cameras, radar and LiDAR) to detect obstructions [1, 3]. However, separation limits the efficiency of human-robot collaboration, while onboard sensors fail in occluded areas such as blind corners [4].

Infrastructure-based solutions using surveillance cameras provide a broader view and reduce blind spots [5, 6], but typically depend solely on semantic classification models [7] for object detection and tracking, which can fail under out-of-distribution conditions. Our approach also uses a network of

surveillance cameras but combines learned semantic models with multi-view 3D reconstruction via space carving, increasing the robustness of detection, 3D positioning, and tracking even of unknown objects (e.g., large loads).

B. Multi-View 3D Reconstruction

Structure from Motion (SfM) and its variants are one of the most popular methods for visual reconstruction of sparse points from multi-view images [8, 9, 10, 11]. However, they rely on high overlap and strong feature matching for triangulation and estimation of 3D points, which are brittle in conditions of low-overlap or highly different viewpoints.

Significant progress was observed in end-to-end learning-based methods [12, 13], achieving dense reconstruction even with low overlap and opposing viewpoints [14]. Despite their success, these approaches suffer from high latency in addition to requiring costly computational resources. To meet the strict real-time requirements of the downstream safety monitoring task, our framework reconstructs 3D scenes from multi-view surveillance cameras and provides dense maps using background subtraction [15] and a model-based shape carving method that projects 3D grid voxels to camera frames, thereby inferring occupied and free areas.

C. 3D Mapping

Reconstructed 3D maps can be represented with raw 3D points, Signed Distance Fields (SDF) [16], 3D voxel grids [17], or probabilistic approaches such as Normal Distribution Transform (NDT) [18] and occupancy maps [19, 20]. Recent occupancy mapping approaches such as Supereight [21] and Wavemap [22] produce multi-resolution voxel maps, with Wavemap achieving state-of-the-art in storage, runtime, and accuracy via wavelet-based compression. However, occupancy estimation typically requires range sensors such as LiDAR, which are prohibitively expensive for infrastructure-mounted large-scale mapping. In contrast, our proposed method leverages the benefits of wavelet decomposition for hierarchical representation and estimates occupancy through shape carving, achieving dense 3D reconstruction using only images.

III. Multi-Resolution Mapping and Safety System

Multi-3DSA is a novel unified framework that enhances the situational awareness in factories by performing dense real-time 3D reconstruction using standard security cameras commonly installed in the infrastructure. At the top level, the system is composed of a *Multi-Resolution 3D Reconstruction and Mapping* module, which reconstructs the scene using multi-view RGB images and represents resulting maps in multiple scales. This module is subdivided into *Background Subtraction*, generating masks for foreground and background, *Shape Carving Reconstruction* to project grid voxels into the masked image space, further estimating occupied and free spaces, and *Wavelet Occupancy Mapping* to compress the map data into a multi-resolution wavelet representation, allowing map queries in coarse and fine resolutions. For scene understanding, 2D semantic segmentation masks inferred from an off-the-shelf trained deep network [23] are projected to 3D data for object tracking and identification. Finally, the tracked 3D objects are used by a *Supervisory System and Management* module that sends safety alerts and stop actions if an accident risk is detected.

A. Multi-resolution 3D Reconstruction and Mapping

This module performs dense 3D reconstruction of scenes using multi-view RGB images from surveillance cameras. Key challenges for our downstream task include: real-time operation, requiring outputs within 500 ms to enable timely responses to accident risks, and scalability to large factory environments.

The reconstruction problem is formulated as a probabilistic occupancy mapping inference. Given a sequence of images $I^k_{0:t}$, where t is the current time and $k \in [0, K]$ denotes the camera index, we aim to build a 3D voxel map $\mathcal{M} = \{m_i\}$ where m_i indicates the i-th voxel binary state (occupied or free). The objective is to estimate the probability of voxels in the scene being occupied or free, as follows:

$$P(m_i|\{I_{0:t}\}^{0:K}). \tag{1}$$

To address the limitations of classical occupancy mapping methods requiring range sensors, we propose a novel probabilistic occupancy estimation with 2D images combining shape carving with occupancy updates.

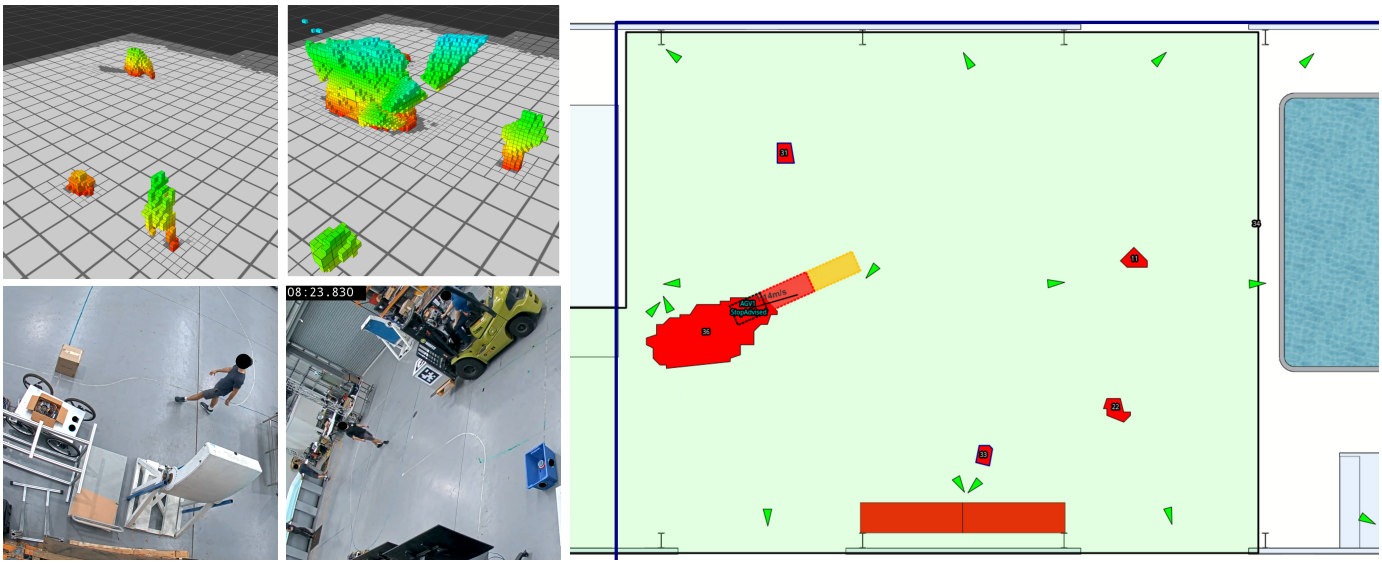
1) Background Subtraction: After receiving each frame from the camera network, Multi-3DSA leverages the background subtraction method EGMM [15] to create 2D masks that separate background and foreground objects.

2) Shape Carving Reconstruction: We reconstruct the scene from the background and foreground masked images using an occupancy-based shape carving technique. Because RGB images do not explicitly provide depth, we estimate the occupancy probabilities by first projecting the 3D voxels to each camera image frame and then evaluating whether the projected voxel area overlaps with foreground or background pixels. The voxels overlapping with foreground masks are updated with a probability of being occupied, whereas voxels projected on background areas are updated with a free occupancy probability. We also define a 3D region of interest within the factory building limits to avoid unnecessary projections.

As required by occupancy mapping methods, our inverse measurement model $P(m_i|I^k_t)$ was adapted for 2D images as follows:

$$P(m_i|I^k_t) = \begin{cases} p_{occ}, \ F^k_t(\pi(\boldsymbol{\nu}_i)) = 1, \\ p_{free}, \ F^k_t(\pi(\boldsymbol{\nu}_i)) = 0, \end{cases} \tag{2}$$

where F^k_t is the foreground mask of I^k_t and $\pi(\boldsymbol{\nu}_i)$ computes the projection of the i-th voxel's axis-aligned bounding box $\boldsymbol{\nu}_i$ to the image space. p_{occ} and p_{free} are fixed occupied and free probability parameters defined by the user based on the measurement confidence.



(a) (b) (c)

Fig. 2: (a) Multi-resolution reconstruction (top) of two people and a box on the floor along with corresponding footage (bottom). (b) Forklift moving towards a worker both captured in the reconstructed map (top) along with the camera view (bottom). Voxels are coloured by height. (c) Web server view of the system with a top-down 2D projection of the reconstructed map, tracked objects (red) with IDs, and detected forklift direction of movement.

The update of a voxel occupancy given camera measurements is then performed in the same fashion as in standard occupancy mapping [20], as follows:

$$\begin{aligned}&\mathcal{L}_{0:t}(P(m_i \mid \{I_{0:t}\}^{0:K})) \\ &= \mathcal{L}_t(P(m_i \mid {I_t}^{0:K})) + \mathcal{L}_{0:t-1}(P(m_i \mid \{I_{0:t-1}\}^{0:K})),\end{aligned} \quad (3)$$

where $\mathcal{L}$ is the log-odds operation $\log \frac{P}{1-P}$.

3) Wavelet Occupancy Mapping: By leveraging a similar approach to [22], we convert the occupancy values to a multi-resolution representation using Haar wavelets as follows:

$$\mathcal{W}(\mathcal{L}_{0:t}) = \mathcal{W}(\mathcal{L}_t) + \mathcal{W}(\mathcal{L}_{0:t-1}), \quad (4)$$

where $\mathcal{W}$ is the wavelet transform applied to the log-odds occupancy. The multi-scale values are then stored in our final map. The wavelets allow us to query the occupancy at different voxel resolutions, which can reduce runtime for coarse voxel operations and storage.

B. Supervisory System and Management

The system has a semantic classification network to detect object classes (such as humans, AGVs or forklifts) and inform the risk of collisions. A view of the supervisory system is illustrated in Fig. 2. When a detected person is inside the risk envelope around a moving object, a safety trigger is actioned on the webview and a safety stop or alert is sent to the moving agent.

IV. Experiments

To evaluate our proposed system, a test bench composed of 22 surveillance Bosch cameras was installed in a workshop building 15 m × 45 m × 5 m in size. The cameras were calibrated for extrinsic and intrinsic parameter estimation. Our method was tested live and also offline on a recorded video stream from the cameras. Two Intel Core Ultra 9 285HX × 24 computers were used during offline tests.

A. Multi-resolution Reconstruction

As observed in Fig. 2 (a) and (b) top, the multi-resolution mapping with shape carving is capable of reconstructing with fine details the shape of larger objects (e.g., human, forklifts) as well as small objects (box), which is valuable for collision risk detection in challenging scenarios (e.g., AGV-human collision). Despite the sparse camera network, the system leverages the multi-view frames to extract the best possible reconstruction. From image inputs to 3D reconstruction, our method takes on average 173 ± 24 ms, meeting the runtime requirement for safety alerting under 500 ms.

V. Conclusion & Future Work

We presented Multi-3DSA, a framework to enhance the safety of humans in mixed autonomy factories. By relying on a sparse network of surveillance cameras, we perform visual 3D reconstruction and multi-resolution mapping using a novel shape carving approach combined with wavelet compression of occupancy estimates. The resulting 3D map can be reconstructed online and in real-time on commercially available compute. Multi-3DSA also performs classification and tracking of reconstructed objects, enabling the detection of accident risks between mobile agents (AGVs, forklifts) and humans. Real-world experiments demonstrated that Multi-3DSA is a cost-effective solution for enabling safety in shared autonomy factories where humans and robots work in the same environment.

VI. Limitations

The proposed system demonstrated good performance in terms of reconstruction quality and processing speed. However, the experimental tests and results also showed limitations that require further research.

The quality of 3D reconstructed maps by visual shape carving is directly related to the number of cameras available. Therefore, more camera viewpoints can offer higher reconstruction quality at the expense of higher costs and computational resources. Due to the limited and sparse camera network in our experiments, which will be the case for any practical deployment, we observed that floating outliers are mistakenly reconstructed since the voxels are not cleared (carved) by any camera view (see Figure 3).

Although floating outliers can be detected from 3D clustering and filtered, heuristic tuning of threshold parameters is required to obtain the best results for each environment. Alternative solutions to reduce outliers while keeping the number of cameras low are the objective of our ongoing and future research. For instance, new 4D radar technologies are cost-effective solutions to provide 3D range measurements and remove outliers if combined with the camera system, but their inherent noise is still a bottleneck for large-scale adoption.

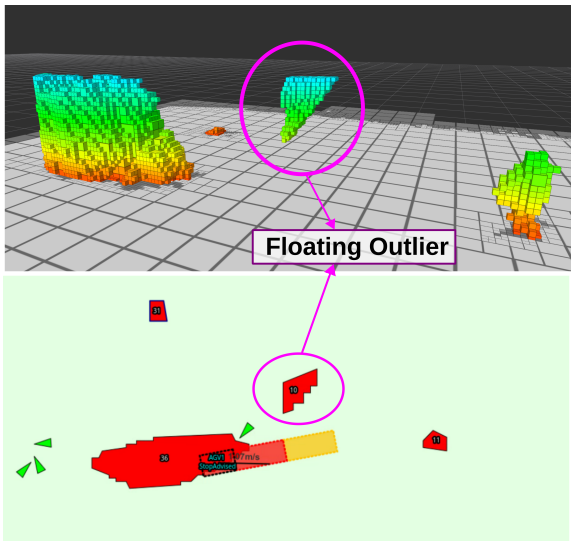


Fig. 3: An example of a floating outlier in the shape carving reconstruction. Such outliers appear in the reconstruction due to the sparse network of camera views which are unable to clear those voxels.

VII. AI usage statement

An AI tool was used only for grammar verification of this text.

References

[1] L. Sabattini, M. Aikio, P. Beinschob, M. Boehning, E. Cardarelli, V. Digani, A. Krengel, M. Magnani, S. Mandici, F. Oleari *et al.*, "The PAN-Robots Project: Advanced Automated Guided Vehicle Systems for Industrial Logistics," *IEEE Robotics & Automation Magazine*, vol. 25, no. 1, pp. 55–64, 2017.

[2] F. Saffre, H. Hildmann, E. Heikkila, T. Malm, and D. Pakkala, "Dynamic and probabilistic safety zones for autonomous mobile robots operating near humans," *Results in Engineering*, vol. 23, p. 102731, 2024.

[3] O. Çatal, S. Leroux, C. De Boom, T. Verbelen, and B. Dhoedt, "Anomaly Detection for Autonomous Guided Vehicles using Bayesian Surprise," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 8148–8153.

[4] C. Drabek, A. Kosmalska, G. Weiss, T. Ishigooka, S. Otsuka, and M. Mizuochi, "Safe Interaction of Automated Forklifts and Humans at Blind Corners in a Warehouse With Infrastructure Sensors," in *International Conference on Computer Safety, Reliability, and Security*. Springer, 2021, pp. 163–177.

[5] E. Cardarelli, L. Sabattini, C. Secchi, and C. Fantuzzi, "Cloud Robotics Paradigm for Enhanced Navigation of Autonomous Vehicles in Real World Industrial Applications," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 4518–4523.

[6] J. Cheng, C. Verhulst, P. De Clercq, S. Van De Velde, S. Sagaert, M. Mertens, M. Birem, M. Deshmukh, N. Broekx, E. Rademakers *et al.*, "Real-Time Warehouse Monitoring with Ceiling Cameras and Digital Twin for Asset Tracking and Scene Analysis," *Logistics*, vol. 9, no. 4, p. 153, 2025.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[8] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski, "Building rome in a day," *Communications of the ACM*, vol. 54, no. 10, pp. 105–112, 2011.

[9] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[10] J. Wang, N. Karaev, C. Rupprecht, and D. Novotny, "VG-GSfM: Visual Geometry Grounded Deep Structure From Motion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 21 686–21 697.

[11] X. Wei, Y. Zhang, Z. Li, Y. Fu, and X. Xue, "Deepsfm: Structure from motion via deep bundle adjustment," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 230–247.

[12] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "DUSt3R: Geometric 3D Vision Made Easy," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 20 697–20 709.

[13] V. Leroy, Y. Cabon, and J. Revaud, "Grounding Image Matching in 3D with MAST3R," in *European conference on computer vision*. Springer, 2024, pp. 71–91.

[14] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "VGGT: Visual Geometry Grounded Transformer," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5294–5306.

[15] S. Liang and D. Baker, "Real-time Background Subtraction under Varying Lighting Conditions," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 9317–9323.

[16] J. Zhang, Y. Yao, and L. Quan, "Learning Signed Distance Field for Multi-view Surface Reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6525–6534.

[17] M. Meilland and A. I. Comport, "On unifying key-frame and voxel-based dense visual SLAM at large scales," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 3677–3683.

[18] P. Manninen, H. Hyyti, V. Kyrki, J. Maanpää, J. Taher, and J. Hyyppä, "Towards High-Definition Maps: A Framework Leveraging Semantic Segmentation to Improve NDT Map Compression and Descriptivity," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 5370–5377.

[19] K. Stepanas, J. Williams, E. Hernández, F. Ruetz, and T. Hines, "OHM: GPU Based Occupancy Map Generation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 078–11 085, 2022.

[20] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: An Efficient Probabilistic 3D Mapping Framework Based on Octrees," *Autonomous robots*, vol. 34, no. 3, pp. 189–206, 2013.

[21] N. Funk, J. Tarrio, S. Papatheodorou, M. Popović, P. F. Alcantarilla, and S. Leutenegger, "Multi-resolution 3d mapping with explicit free space representation for fast and accurate mobile robot motion planning," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3553–3560, 2021.

[22] V. Reijgwart, C. Cadena, R. Siegwart, and L. Ott, "Efficient volumetric mapping of multi-scale environments using wavelet-based compression," 2023-07.

[23] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017. [Online]. Available: https://arxiv.org/abs/1706.05587