Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 2, April 2026, pp. 1106~1113
ISSN: 2302-9285, DOI: 10.11591/eei.v15i2.9765

Real-time object detection and XAI-based activation map visualization using YOLOv8s

Ashaari Yusof[1], Muhammad Hishamuddin[1], Md Jakir Hossen[2]
[1]Centre for Robotics and Sensing Technologies, Telekom Malaysia Research and Development, Cyberjaya, Malaysia
[2]Faculty of Engineering and Technology, Multimedia University, Melaka, Malaysia

Article Info

Article history:

Received Dec 23, 2024
Revised Sep 9, 2025
Accepted Feb 22, 2026

Keywords:

Activation map visualization
Explainable artificial intelligence
Deep learning interpretability
Real-time object detection
YOLOv8

ABSTRACT

This study introduced a methodology for real-time object detection and interpretability using YOLOv8s, trained on the MS common objects in context (COCO) dataset. The system captured live webcam footage, processes frames resized to 640×384, and applies YOLOv8s to detect objects with bounding boxes, labels, and confidence scores. YOLOv8s architecture comprising a CSPDarknet53-based backbone, neck, and head ensures efficient feature extraction and accurate detection. To enhance model transparency, activation map generation is implemented by attaching forward hooks to intermediate convolutional layers. Feature maps are captured during the forward pass, averaged, normalized, and resized to match the original image dimensions. This visualization highlights regions influencing the model's predictions, aligning with explainable artificial intelligence (XAI) principles. Experimental results demonstrate high detection accuracy and effective interpretability in indoor environments, making the framework suitable for robotics applications requiring both precision and transparency. The proposed method offers a practical and explainable solution for real-time scene understanding in intelligent systems.

Corresponding Author:

Ashaari Yusof
Centre for Robotics and Sensing Technologies, Telekom Malaysia Research and Development
Cyberjaya, Malaysia
Email: ashaari@tmrnd.com.my

1. INTRODUCTION

The robotics industry is at the forefront of the automation revolution, hugely driven by advancements in artificial intelligence (AI), machine learning (ML), and computer vision technologies. Mobile robots deployment such as autonomous mobile robots (AMRs) in indoor environments is becoming increasingly prevalent. These indoor robots are designed to operate independently, utilizing sophisticated sensors and AI algorithms to navigate complex spaces and perform a variety of tasks such as housekeeping and patient care [1]-[3]. To accommodate these growing number of robotics applications, the robots need to better perceive their surroundings to achieve intelligent operation.

A popular area of research is to train these robots to interact with their surroundings and people by leveraging their image recognition capabilities, a primary focus of computer vision [4], [5]. This enables them to see, comprehend, and independently adapt to their natural surroundings [6]. The foundation of their decision-making and behavior optimization in a complex indoor environment is scene understanding, which is also essential in evaluating how well the indoor robots operate [7], [8]. The study of object detection holds substantial practical significance, as scene understanding is inherently linked to the identification and comprehension of key objects within the environment.

Additionally, one of the more intricate challenges in computer vision is identifying indoor objects and scenes. Accurately determining the location, size, and category of objects is essential for indoor scene object detection technology, as it significantly enhances the precision of research on indoor mobile robot positioning and navigation [9], [10]. Prior to the development of deep learning technology, conventional object detection techniques primarily relied on manually created features, such histograms of oriented gradient (HOG) [11], scale-invariant feature transform (SIFT) [12], and haar-like feature (HAAR) [13]. The methods of hand design features have issues with low detection accuracy, poor generalization, and poor robustness because of the imaging environment and the object itself [14]. The traditional methods, which rely on manual design elements, perform poorly in complicated indoor environments and struggle to satisfy people's need for high-performance object recognition.

To address the challenges of accurate object recognition for indoor navigation, the primary focus of this paper is to explore a novel methodology for real-time object detection and enhanced interpretability using YOLOv8. The main contributions of this paper are as:

- Implementation of YOLOv8s to detect objects in live webcam footage with high accuracy and confidence scores.
- Real-time frame capture from webcams using a JavaScript-based approach, with consistent resizing for optimal model input.
- Generation of activation map heatmaps to interpret the model's focus areas, aligning with explainable artificial intelligence (XAI) principles.

2. RELATED WORK

There exist a large body of work that underscores significant advancements in object detection methodologies for robotics and other related applications. Singh *et al.* [15] and Chen and Li [16] both focus on enhancing navigation capabilities for AMRs in indoor environments through advanced object detection methods. Singh *et al.* [15] introduced a modified you only look once (YOLO) algorithm that improved object detection and recognition efficiency, demonstrating superior performance metrics such as mean average precision (mAP) and reduced inference time compared to traditional methods. This approach is particularly relevant for service robots operating in constrained indoor spaces. Similarly, Chen and Li [16] proposed ATopNet, a robust visual localization model that incorporates pose and speed enhancements, achieving real-time localization with significant accuracy. Both studies highlight the importance of optimizing algorithms for low-computing devices, which is crucial for deploying effective object detection systems in practical indoor AMR applications.

The integration of multimodal sensing techniques in AMR is explored by Chen *et al.* [17] and Wunderle *et al.* [18]. Chen *et al.* [17] developed a system that combines RGBD images with 2D light detection and ranging (LIDAR) data to improve movement accuracy and positioning in indoor environments, achieving impressive localization metrics such as 98.3% accuracy. Wunderle *et al.* [18] emphasized the need for comprehensive system architectures to enhance safety and reliability in AMR operations, identifying critical gaps between current safety standards and state-of-the-art pedestrian detection systems. Their work underscores the necessity for datasets tailored to train neural networks for safety applications.

Moreover, Aulia *et al.* [19] and Cherubin *et al.* [20] further illustrate the ongoing advancements in object detection systems tailored for AMRs operating in diverse environments. Aulia *et al.* [19] developed a new CNN-based object detection system utilizing real-world vehicle datasets, achieving high classification accuracy across varying lighting conditions that contributes to reliable operation in urban settings. Cherubin *et al.* [20] exploration of YOLO object detection on low-cost mobile robots emphasizes affordability without compromising performance quality, making advanced technologies accessible for practical deployment.

Also, within the YOLO framework, Brand *et al.* [21] and Yang *et al.* [22] contribute to the discussion on enhancing object detection capabilities through innovative methodologies that leverage YOLO architectures. Brand *et al.* [21] explored ultrasonic object detection methods combined with YOLOv8 for outdoor applications, achieving a mAP of 0.685 by converting ultrasonic data into images for classification tasks. This novel approach demonstrates the versatility of YOLOv8 beyond traditional visual inputs. On the other hand, Yang *et al.* [22] introduced a dynamic visual simultaneous localization and mapping (SLAM) algorithm based on lightweight YOLOv8, which improves positioning accuracy in high-dynamic scenes by filtering out dynamic objects that could disrupt SLAM algorithms. Together, these studies showcase the adaptability of YOLO models across different sensing modalities and environments.

Lastly, the application of XAI techniques in conjunction with YOLO models is highlighted by Dewangan and Gupta [23] and Liu *et al.* [24]. Dewangan and Gupta [23] evaluated a YOLOv8-based framework for indoor fire and smoke detection, utilizing XAI methods like LIME and SHAP to enhance interpretability alongside performance metrics such as mAP and F1 score. Their findings indicate that while YOLOv8n achieved superior mAP scores, challenges remain in reliably detecting smoke. Similarly,

Liu *et al*. [24] work on an improved YOLOv5 model for safety belt detection emphasizes the integration of advanced techniques to enhance accuracy while reducing computational costs, demonstrating the importance of explainability in safety related applications.

3. METHOD

This study employs YOLOv8s for object detection within captured frames. The architecture of the YOLOv8 is shown in Figure 1. The components of this model are as follows. The backbone extracts important features from the input image using cross stage partial networks (CSPs). The bottleneck CSP module is responsible for feature extraction, reducing redundancy efficiently during optimization. The spatial pyramid pooling (SPP) module expands the network's receptive field and captures elements of various scales. The backbone is based on the CSPDarknet53 architecture. The second component, the neck, is an intermediate component that connects the backbone to the head. It aggregates and refines the features extracted by the backbone, often focusing on enhancing the spatial and semantic information across different scales. The neck may include additional convolutional layers, feature pyramid networks (FPN), or other mechanisms to improve the representation of the features. The last component is the head. Its main responsibility is to supervise the last stage of detection. The final output vectors, which include bounding box information, object absence scores, and class probabilities, are produced by utilizing anchor boxes.

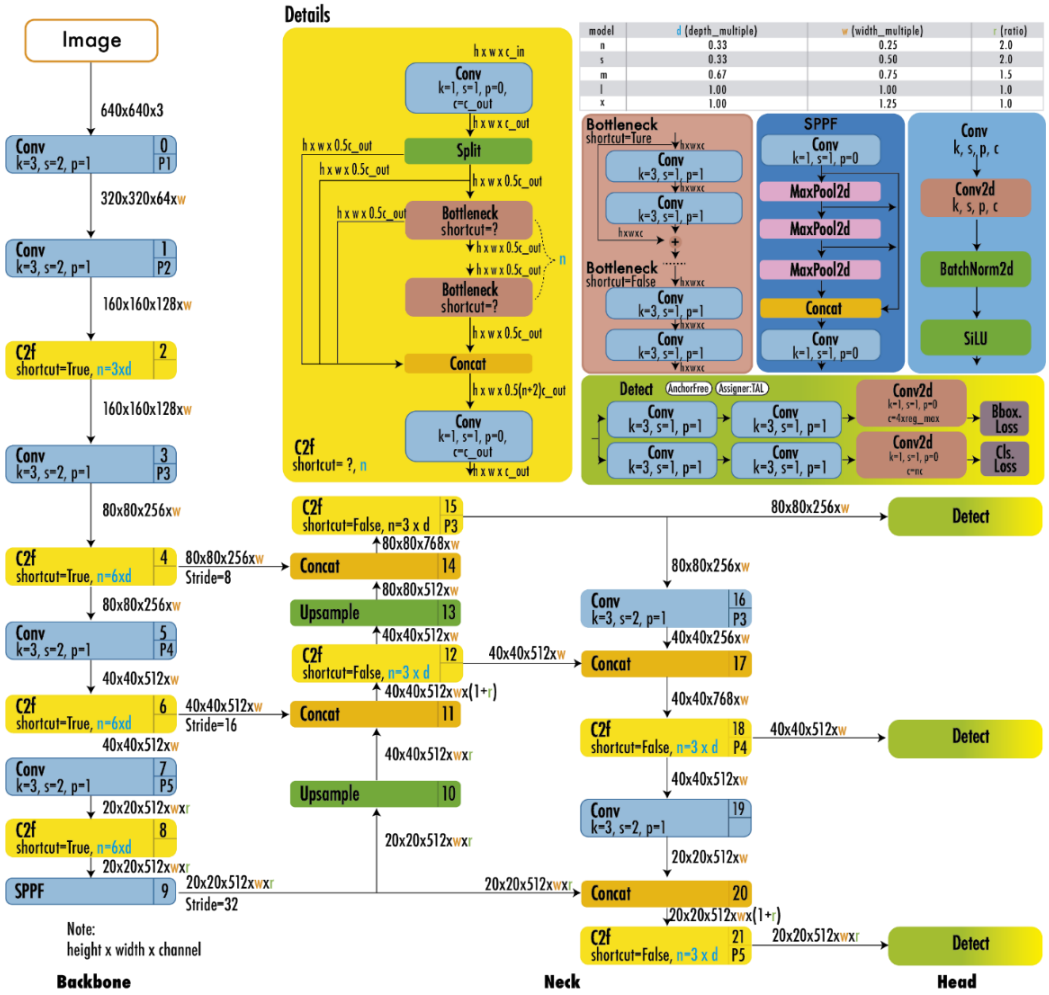

Figure 1. Architecture of YOLOv8 [25]

In this work, the process of object detection using the YOLOv8s model begins with the training phase, where the model is trained on common objects in context (COCO) dataset [26] which serves as the

foundational data for the training process. This training involves feeding the YOLOv8 architecture with images from the COCO dataset. The backbone utilizes CSPs to extract critical features while minimizing redundancy. The BottleneckCSP module enhances feature extraction efficiency, and the SPP module increases the receptive field to capture multi-scale elements. The neck aggregates and refines these features, enhancing spatial and semantic information before passing them to the Head. The Head generates final output vectors that include bounding box coordinates, object confidence scores, and class probabilities based on anchor boxes.

Figure 2 illustrates the workflow diagram of this study, focussing on real time object detection and activation map visualisation via YOLOv8s. The trained YOLOv8 model is employed for real-time object detection, initiated by capturing frames using a webcam. In pre-processing, these frames are decoded into a NumPy array, resized to (640 and 384), and passed through the YOLO model to extract bounding boxes, confidences, and class IDs. The annotated frames are saved and displayed in a continuous loop until interrupted by the user.

For interpretability, the methodology incorporates the generation of activation maps. Each saved image is passed through the model to extract and normalize feature maps, which are then used to create a heatmap overlay. This overlay is displayed alongside the original image, providing a visual representation of the model's focus areas. This is achieved by registering a forward hook to capture intermediate layer activations, thereby enhancing the transparency and explainability of the object detection process. This systematic approach ensures both the accuracy of object detection and the interpretability of the model's decisions, which is crucial for applications requiring transparency and trust such as the ones associated with indoor robotics.

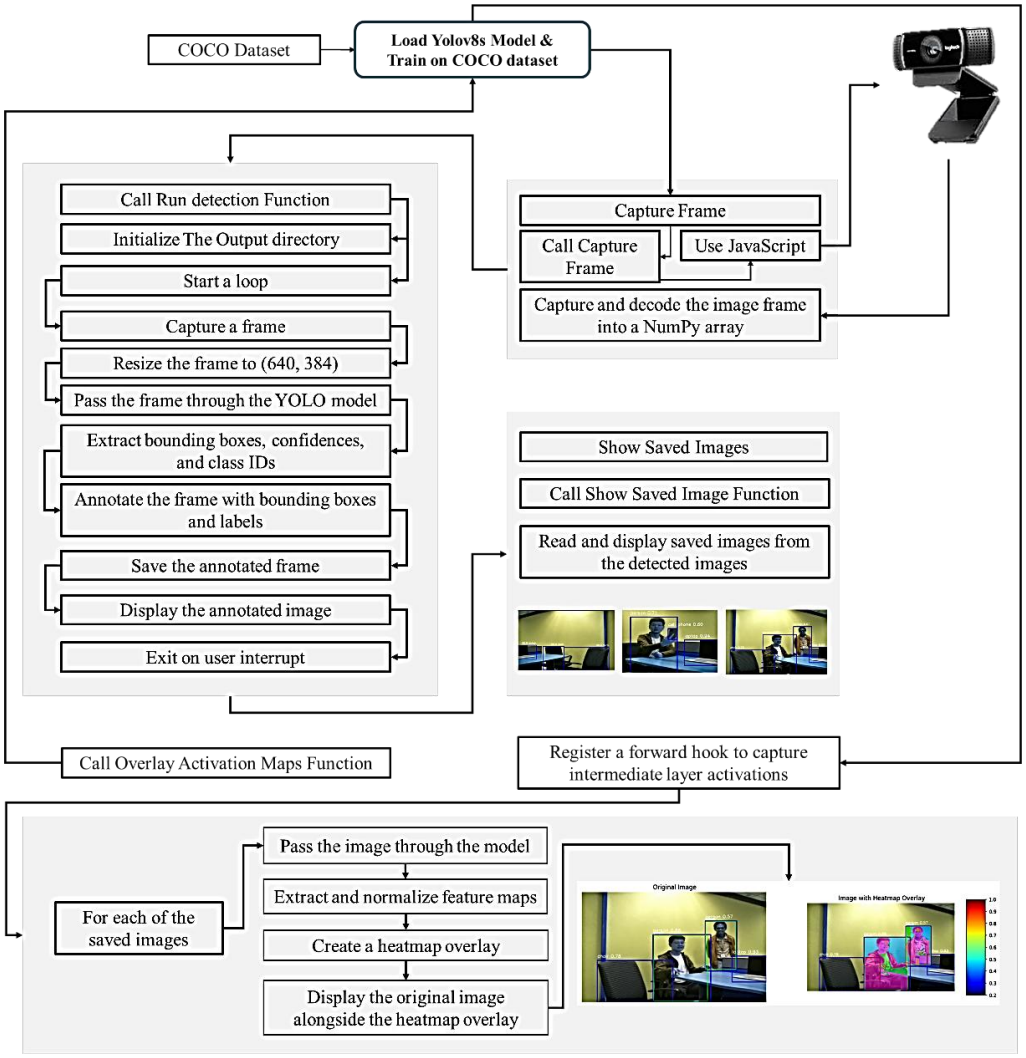

Figure 2. YOLOv8s model training, frame capture, and activation maps workflow

4. RESULTS AND DISCUSSION

4.1. Real time object detection

Figure 3 demonstrates real-time object detection results using the YOLOv8s model in an office setting, consisting three sequential images among 779 images showcasing various detections, each marked with bounding boxes and labeled with the object name and confidence score. In the first image, two chairs are detected with confidence scores of 0.83 and 0.85, focusing on an empty section of the office. The second image captures a person detected with a confidence score of 0.71, along with a laptop and a cell phone detected with confidence scores of 0.94 and 0.60, respectively. The third image provides a broader view of the office, detecting two persons with confidence scores of 0.85 and 0.57, alongside a chair detected at 0.78 and laptop detected at 0.93. The figures effectively illustrate YOLOv8s' ability to identify multiple objects in an office environment in real time with varying confidence levels.

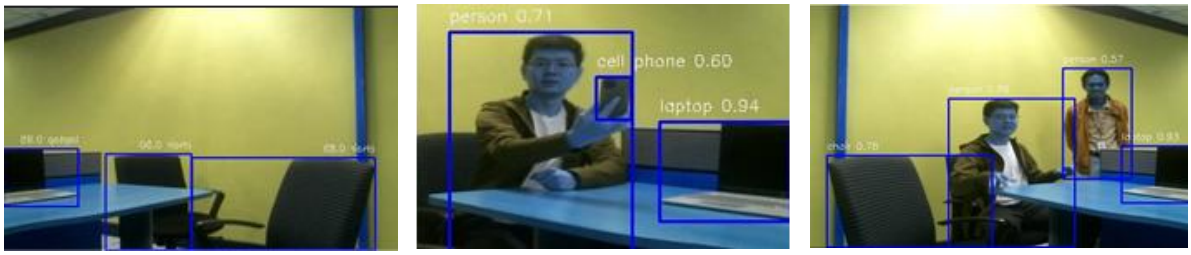

Figure 3. Real time object detection images using YOLOv8s

4.3. Activation map

Figure 4 consists two visuals demonstrating object detection and model interpretability through activation maps. Figure 4(a) shows two persons detected within an indoor environment, highlighted by a blue bounding box. The background features include walls and a structured ceiling with lighting, but the model focuses solely on the persons, chair and laptop as the objects of interest. Figure 4(b) shows the heatmap overlay image of 'persons' object class type, providing an interpretability layer by displaying a heatmap superimposed on the original image. The heatmap, created using the JET colormap, employs warm colors red, green, blue (RGB) to highlight areas the model considers significant for object detection. Together, these visuals highlight the effectiveness of the object detection model and its explainability, offering insights into how the model identifies objects and the specific features it prioritizes during decision-making.

(a) (b)

Figure 4. Object detection and activation map visualization using YOLOv8sm: (a) original image showing detected objects (persons, chair, and laptop) with bounding boxes and (b) corresponding activation map overlay using JET colormap, highlighting regions of interest influencing the model's predictions

This dual focus on real-time detection and interpretability offers a balanced approach to both practical application and model transparency. While the YOLOv8s model demonstrates strong detection capabilities, the addition of activation maps addresses a common limitation in deep learning models: the lack of insight into how predictions are made. By visualizing focus areas, the methodology fosters trust and usability, making it particularly valuable in applications where understanding model behavior is crucial. The

proposed framework, therefore, not only achieves high detection accuracy but also provides stakeholders with actionable insights into the inner workings of the model.

Although the images in Figure 4 show promising results, one key aspect of the experiment that could be further improved is the quality of the captured frame image. The 'chair' and 'laptop' object classes were not properly identified in Figure 4(b) due to the average image quality. Capturing video images in a well-lit indoor environment, combined with a higher quality webcam, could enhance heatmap detection during the overlay process. Additionally, incorporating image pre-processing techniques could further improve the overall quality of the captured frames.

5. CONCLUSION

This study successfully demonstrates the integration of YOLOv8s for real-time object detection and interpretability through activation map visualization, offering a dual benefit of high detection accuracy and model interpretability. By leveraging YOLOv8s' robust architecture and implementing activation maps through intermediate layer hooks, the framework provides clear insights into the model's decision-making process. The results validate the system's effectiveness in indoor environments, demonstrating its potential for deployment in robotics and smart surveillance applications. Importantly, the inclusion of interpretability features addresses a critical gap in deep learning systems—transparency and trust. This empowers stakeholders to better understand and validate AI-driven decisions. Future work may focus on improving image quality and exploring additional XAI techniques to further enhance model explainability and robustness. Overall, the proposed framework contributes meaningfully to the advancement of intelligent, transparent, and reliable object detection systems.

FUNDING INFORMATION

The authors thank Telekom Malaysia for their funding under RDTC241131 project.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ashaari Yusof	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Muhammad Hishamuddin	✓	✓		✓	✓	✓	✓	✓		✓	✓			
Md Jakir Hossen	✓	✓		✓						✓		✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] G. Wang, Y. Hu, X. Wu, and H. Wang, "Residual 3-D Scene Flow Learning with Context-Aware Feature Extraction," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–9, 2022, doi: 10.1109/TIM.2022.3166147.

[2] S. Liu, G. Tian, Y. Zhang, M. Zhang, and S. Liu, "Service planning oriented efficient object search: A knowledge-based framework for home service robot," *Expert Syst. Appl.*, vol. 187, p. 115853, Jan. 2022, doi: 10.1016/j.eswa.2021.115853.

[3] K. Wang, X. Li, J. Yang, J. Wu, and R. Li, "Temporal action detection based on two-stream You Only Look Once network for elderly care service robot," *Int. J. Adv. Robot. Syst.*, vol. 18, no. 4, Jul. 2021, doi: 10.1177/17298814211038342.

[4] R. Ma, Z. Zhang, and E. Chen, "Human Motion Gesture Recognition Based on Computer Vision," *Complexity*, vol. 2021, no. 1, Jan. 2021, doi: 10.1155/2021/6679746.
[5] S. Cui, R. Wang, J. Hu, C. Zhang, L. Chen, and S. Wang, "Self-Supervised Contact Geometry Learning by GelStereo Visuotactile Sensing," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–9, 2022, doi: 10.1109/TIM.2021.3136181.
[6] Z. Huang, C. Lv, Y. Xing, and J. Wu, "Multi-Modal Sensor Fusion-Based Deep Neural Network for End-to-End Autonomous Driving with Scene Understanding," *IEEE Sens. J.*, vol. 21, no. 10, pp. 11781–11790, May 2021, doi: 10.1109/JSEN.2020.3003121.
[7] Y. Cheng, Y. Yang, H. B. Chen, N. Wong, and H. Yu, "S3-Net: A Fast Scene Understanding Network by Single-Shot Segmentation for Autonomous Driving," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 5, pp. 1–19, Oct. 2021, doi: 10.1145/3470660.
[8] J. Fan, P. Zheng, and S. Li, "Vision-based holistic scene understanding towards proactive human–robot collaboration," *Robot. Comput. Integr. Manuf.*, vol. 75, p. 102304, Jun. 2022, doi: 10.1016/j.rcim.2021.102304.
[9] X. Zhao, T. Zuo, and X. Hu, "OFM-SLAM: A Visual Semantic SLAM for Dynamic Indoor Environments," *Math. Probl. Eng.*, vol. 2021, pp. 1–16, Apr. 2021, doi: 10.1155/2021/5538840.
[10] M. Y. Moemen, H. Elghamrawy, S. N. Givigi, and A. Noureldin, "3-D Reconstruction and Measurement System Based on Multimobile Robot Machine Vision," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021, doi: 10.1109/TIM.2020.3026719.
[11] R. J. A. Ali and A. Almamori, "Human Action Recognition Based on Histograms of Oriented Gradients from RGBD Cameras," in *AIP Conference Proceedings*, 2024, p. 040003, doi: 10.1063/5.0207137.
[12] G. Tang, Z. Liu, and J. Xiong, "Distinctive image features from illumination and scale invariant keypoints," *Multimed. Tools Appl.*, vol. 78, no. 16, pp. 23415–23442, Aug. 2019, doi: 10.1007/s11042-019-7566-8.
[13] K. Shaheed, Q. Abbas, and M. Kumar, "Automatic diagnosis of CoV-19 in CXR images using haar-like feature and XgBoost classifier," *Multimed. Tools Appl.*, vol. 83, no. 26, pp. 67723–67745, Jan. 2024, doi: 10.1007/s11042-024-18330-9.
[14] T. Sharma *et al.*, "Deep Learning-Based Object Detection and Classification for Autonomous Vehicles in Different Weather Scenarios of Quebec, Canada," *IEEE Access*, vol. 12, pp. 13648–13662, 2024, doi: 10.1109/ACCESS.2024.3354076.
[15] K. J. Singh, D. S. Kapoor, K. Thakur, A. Sharma, and X. Z. Gao, "Computer-Vision Based Object Detection and Recognition for Service Robot in Indoor Environment," *Comput. Mater. Contin.*, vol. 72, no. 1, pp. 197–213, 2022, doi: 10.32604/cmc.2022.022989.
[16] H. H. Chen and C. H. G. Li, "ATopNet: Robust Visual Localization for AMR Navigation," in *Proc. 2022 IEEE/ASME Int. Conf. Adv. Intell. Mechatronics (AIM)*, Sapporo, Japan, 2022, pp. 275-281, doi: 10.1109/AIM52237.2022.9863347.
[17] C. W. Chen, C. L. Lin, J. J. Hsu, S. P. Tseng, and J. F. Wang, "Design and Implementation of AMR Robot Based on RGBD, VSLAM and SLAM," in *Proc. 2021 9th Int. Conf. Orange Technol. (ICOT)*, Tainan, Taiwan, 2021, pp. 1-5, doi: 10.1109/ICOT54518.2021.9680621.
[18] Y. Wunderle, E. Lyczkowski, and S. Hohmann, "Safe object detection in AMRs - a Survey," in *Proc. 2024 IEEE 33rd Int. Symp. Ind. Electron. (ISIE)*, Ulsan, Korea, Republic of, 2024, pp. 1-8, doi: 10.1109/ISIE54533.2024.10595686.
[19] U. Aulia, I. Hasanuddin, M. Dirhamsyah, and N. Nasaruddin, "A new CNN-BASED object detection system for autonomous mobile robots based on real-world vehicle datasets," *Heliyon*, vol. 10, no. 15, p. e35247, Aug. 2024, doi: 10.1016/j.heliyon.2024.e35247.
[20] S. Cherubin, W. Kaczmarek, and M. Siwek, "YOLO object detection and classification using low-cost mobile robot," *Prz. Elektrotechniczny*, vol. 1, no. 9, pp. 29–33, Sep. 2024, doi: 10.15199/48.2024.09.04.
[21] L. Brand, Y. Wunderle, and S. Hohmann, "Ultrasonic Object Detection and Classification for AMR Safety," in *Proc. 2024 IEEE 29th Int. Conf. Emerg. Technol. Factory Autom. (ETFA)*, Padova, Italy, 2024, pp. 01-06, doi: 10.1109/ETFA61755.2024.10711122.
[22] Z. Yang, H. Zhang, and X. Fan, "Dynamic Visual SLAM Algorithm Based on Lightweight YOLOv8," in *Proc. 2024 3rd Int. Conf. Artif. Intell. Comput. Inf. Technol. (AICIT)*, Yichang, China, 2024, pp. 1-4, doi: 10.1109/AICIT62434.2024.10729997.
[23] D. K. Dewangan and G. P. Gupta, "Explainable AI and YOLOv8-based Framework for Indoor Fire and Smoke Detection," in *Proc. 2024 IEEE Int. Conf. Inf. Technol., Electron., Intell. Commun. Syst. (ICITEICS)*, Bangalore, India, 2024, pp. 1-6, doi: 10.1109/ICITEICS61368.2024.10624874.
[24] L. Liu, K. Huang, Y. Bai, Q. Zhang, and Y. Li, "Real-time detection model of electrical work safety belt based on lightweight improved YOLOv5," *J. Real-Time Image Process.*, vol. 21, no. 4, p. 151, Aug. 2024, doi: 10.1007/s11554-024-01533-6.
[25] J. Terven, D. M. Córdova-Esparza, and J. A. Romero-González, "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Mach. Learn. Knowl. Extr.*, vol. 5, no. 4, pp. 1680–1716, Nov. 2023, doi: 10.3390/make5040083.
[26] D. Shah, "COCO Dataset: All You Need to Know to Get Started,". [Online]. Available: https://www.v7labs.com/blog/coco-dataset-guide, 2023. (Accessed: Nov. 19, 2024).

BIOGRAPHIES OF AUTHORS

Ashaari Yusof is from the Centre for Robotics and Sensing Technologies, TM Research and Development Sdn Bhd, Malaysia. He is currently developing smart use cases focusing on robotic applications and urban forestry solutions. His research interests include IoT intelligent services and computer vision. He is a senior researcher with 20+ years of academia-industry collaborative experience. He can be contacted at email: ashaari@tmrnd.com.my.

Muhammad Hishamuddin is a researcher at the Centre for Robotics and Sensing Technologies, TM Research and Development Sdn Bhd, Malaysia, specializing in IoT, robotics, UAVs, and industrial automation. With over 16 years of experience, he is skilled in ROS, Python/C++ programming, mechanical and electrical system design, and manufacturing technologies. A graduate of Pennsylvania State University with a degree in Agricultural & Biological Engineering (2008), he focuses on advancing robotics and IoT innovations. He can be contacted at email: muhammad@tmrnd.com.my.

Md Jakir Hossen is currently working as an Associate Professor in the Department of Robotics and Automation, Faculty of Engineering and Technology, Multimedia University, Malaysia. He received the master degree in communication and network engineering from Universiti Putra Malaysia, Malaysia in 2003. He received the Ph.D. degree in smart technology and robotic engineering from Universiti Putra Malaysia, Malaysia in 2012. His research interests are application of artificial intelligence techniques in data analytics, robotics control, data classifications, and predictions. He can be contacted at email: jakir.hossen@mmu.edu.my.