Automatic Detection of Personal Protective Equipment in Industrial Environments Using YOLO and the NAO Humanoid Robot

Detecção Automática de Equipamentos de Proteção Individual em Ambientes Industriais Utilizando YOLO e o Robô Humanoide NAO

ABSTRACT

Occupational safety in industrial environments represents a critical challenge that requires robust and efficient technological solutions. This work presents the development and implementation of a proof-of-concept system for detecting Personal Protective Equipment (PPE) using the You Only Look Once (YOLO) algorithm integrated with the NAO humanoid robot. The model was trained via fine-tuning on a specialized dataset containing annotations for different PPE classes, including safety helmets, reflective vests, safety boots, gloves, and masks. System validation was performed on five synthetic images generated by Artificial Intelligence under controlled conditions. Model-level inference times ranged from 0.01 to 0.04 seconds, while the implemented application executed detection cycles every 2 seconds. The obtained results show confidence rates exceeding 85% for the main PPE classes, notably for Safety Vest (90.05%), Person (87.93%), and Hardhat (86.17%). Integration with the NAO robot demonstrates the feasibility of combining robotic visual sensing with offboard deep-learning inference for occupational-safety monitoring. The study therefore establishes a proof-of-concept basis for future validation in real industrial environments. The analysis focused on three main dimensions: class confidence, processing latency, and detection coverage across controlled synthetic scenes.

Keywords

Object Detection; YOLO; Personal Protective Equipment; Humanoid Robotics; Occupational Safety; Computer Vision.

RESUMO

A segurança ocupacional em ambientes industriais representa um desafio crítico que exige soluções tecnológicas robustas e eficientes. Este trabalho apresenta o desenvolvimento e a implementação de um sistema de prova de conceito para detecção de Equipamentos de Proteção Individual (EPIs) utilizando o algoritmo You Only Look Once (YOLO) integrado ao robô humanoide NAO. O modelo foi treinado por meio de fine-tuning em um conjunto de dados especializado, contendo anotações para diferentes classes de EPIs, incluindo capacetes de segurança, coletes refletivos, botas de segurança, luvas e máscaras. A validação do sistema foi realizada com cinco imagens sintéticas geradas por Inteligência Artificial em condições controladas. Os tempos de inferência do modelo variaram entre 0,01 e 0,04 segundos, enquanto a aplicação implementada executou ciclos de detecção a cada 2 segundos. Os resultados obtidos apresentam taxas de confiança superiores a 85% para as principais classes de EPIs, com destaque para Safety Vest (90,05%), Person (87,93%) e Hardhat (86,17%). A integração com o robô NAO demonstra a viabilidade de combinar sensoriamento visual robótico com inferência de aprendizado profundo executada externamente para o monitoramento da segurança ocupacional. O estudo estabelece, assim, uma base de prova de conceito para futura validação em ambientes industriais reais. A análise concentrou-se em três dimensões principais: confiança por classe, latência de processamento e cobertura de detecção nas cenas sintéticas controladas.

Palavras-Chave

Detecção de Objetos; YOLO; Equipamentos de Proteção Individual; Robótica Humanoide; Segurança do Trabalho; Visão Computacional.

1. INTRODUCTION

Occupational safety constitutes one of the primary concerns in industrial environments, particularly in sectors such as construction, mining, petrochemicals, and heavy manufacturing. In these settings, workers are constantly exposed to significant physical risks [1]. According to the International Labour Organization, approximately 2.78 million workers die annually due to work-related accidents and occupational diseases. Additionally, 374 million workers suffer non-fatal injuries [2].

In this context, the proper use of Personal Protective Equipment (PPE) represents a critical line of defense against accidents and injuries. Its use is regulated by stringent standards in nearly all industrialized countries. However, manual monitoring of PPE compliance presents notable limitations, including human capacity for continuous supervision, inspector fatigue, and the impossibility of covering large operational areas [3].

Recent advances in deep learning techniques, particularly convolutional neural network-based object detection, have enabled the development of automated systems capable of intelligent visual monitoring with high accuracy and speed [4]. The *You Only Look Once* (YOLO) algorithm has emerged as one of the most efficient architectures for real-time object detection, offering a favorable balance between accuracy and computational speed [5].

Concurrently, humanoid robotics has made remarkable progress. Platforms such as NAO demonstrate increasing capabilities for interacting with human environments and performing complex perception and navigation tasks [6]. Unlike fixed cameras, mobile platforms such as NAO can navigate around obstacles, reposition dynamically, and interact verbally with workers, enabling coverage of areas inaccessible to static surveillance systems.

The convergence of these technologies opens unprecedented opportunities for autonomous safety monitoring systems. Such systems can navigate industrial facilities, identify safety protocol violations, and issue real-time alerts.

This work presents the development of an integrated system combining YOLO-based PPE detection with the NAO robotic platform. This integration provides a basis for robotic safety-compliance monitoring and future mobile inspection applications. The system was developed through fine-tuning a pre-trained YOLO model using a specialized PPE dataset from the RoboFlow platform [7]. Validation was then performed on synthetic images generated by artificial intelligence to assess robustness and generalization.

The proposed methodology contributes to advances in occupational safety automation. It provides a technically feasible proof-of-concept architecture for subsequent validation in real industrial environments. The results demonstrate that the system can detect multiple PPE classes simultaneously with high confidence, while maintaining sufficiently low latency for real-time applications. These results motivate further evaluation of the proposed approach as a tool for continuous automated occupational-safety monitoring.

The remainder of this paper is organized as follows. Section 2 presents the conceptual grounding related to object detection, YOLO, humanoid robotics, and PPE monitoring. Section 3 describes the study design, including dataset preparation, model fine-tuning, integration with the NAO robot, and the validation procedure. Section 4 introduces the mathematical formulation of the proposed detection process, while Section 5 presents the implemented algorithm in pseudocode form. Section 6 discusses the experimental results obtained from the synthetic test scenarios, including detection confidence and inference time. Finally, Section 7 presents the main conclusions, limitations, and directions for future work.

2. CONCEPTUAL GROUNDING

Automatic object detection in digital images represents one of the fundamental problems in computer vision. Its applications range from autonomous vehicles to intelligent surveillance systems [8]. Traditional approaches relied on handcrafted feature descriptors, such as Histogram of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT), combined with classifiers like Support Vector Machines. However, these methods presented limitations in terms of generalization and robustness to variations in lighting, occlusion, and perspective [9].

The YOLO algorithm, introduced by Redmon *et al.* in 2016, revolutionized real-time object detection by reformulating the problem as a direct regression task of bounding box coordinates and class probabilities [4]. The original architecture divided the input image into a regular grid and assigned each cell the responsibility to detect objects whose centers fell within that cell. This enabled extremely fast processing, capable of handling images at over 45 frames per second on conventional hardware. Subsequent YOLO versions introduced additional improvements. YOLOv2 incorporated batch normalization, anchor boxes, and a deeper Darknet-19 architecture, while YOLOv3 adopted multi-scale predictions and residual connections [10], [11]. The evolution continued with YOLOv4 and YOLOv5, which introduced advanced augmentation techniques, architectural optimizations, and training enhancements. These versions established YOLO as state-of-the-art for real-time object detection [5].

NAO has been employed in various human-robot interaction applications, such as assisting children with autism, museum guides, and receptionists in corporate environments [12]. Its software architecture is based on the NAOqi framework, which provides high-level abstraction for motor control, sensory processing, and communication. This facilitates the development of complex applications through APIs in Python, C++, and other languages [13].

The integration of advanced computer vision algorithms with mobile robotic platforms such as NAO enables the development of autonomous systems. These systems can navigate industrial environments, interact naturally with workers, and perform continuous monitoring of safety protocols without constant human intervention. Pre-trained YOLO models on the COCO dataset, containing over 200,000 annotated images with 80 common object classes, provide robust visual representations. These representations can be adapted to specialized domains via fine-tuning on smaller, task-specific datasets [14].

The generation of synthetic data using generative AI models has recently emerged as a promising technique to augment training datasets and assess model robustness [15]. Generative adversarial and latent diffusion models, as implemented in systems such as DALL-E, MidJourney, and Grok, have shown increasing capacity to produce high-

quality photorealistic images from textual descriptions. This approach enables the rapid creation of diverse test scenarios without the need for physical image capture [16].

Using synthetic images for PPE detection validation constitutes an innovative strategy. It allows evaluating model robustness against variations in lighting, poses, occlusions, and other confounding factors in a controlled and scalable manner. Figure 1 illustrates the setup designed for the NAO robot to view these images via a monitor and subsequently classify their content automatically.



Figure 1: Controlled proof-of-concept setup used for synthetic-scene evaluation.

This work integrates these technological streams by combining YOLO fine-tuning on a specialized PPE dataset, implementation on the NAO humanoid robotic platform, and validation using AI-generated synthetic images. This integration provides a proof-of-concept framework for robotic occupational-safety monitoring and a basis for subsequent validation under representative industrial conditions.

3. STUDY DESIGN

The development of the proposed system was structured into four main stages: acquisition and preparation of the training dataset, YOLO model fine-tuning, integration with the NAO robotic platform, and validation through inference on synthetic images. The training data used in this study were obtained through the Roboflow platform [7], a repository and tooling ecosystem for annotated computer-vision datasets [17]. The local training export employed by the experiment used the application labels Hardhat, Person, Safety_Boots, Safety_Gloves, Safety_Mask, and Safety_Vest. The image counts associated with this local experimental set are presented in Table 1. Because hosted datasets can evolve across versions, the values reported here refer specifically to the training material used to produce the evaluated `best.pt` weights.

Table 1: Number of images per class

Class	Images
Hardhat	354
Person	287
Safety_Boots	218
Safety_Gloves	313
Safety_Mask	250
Safety_Vest	250

Each image in the dataset was previously annotated with YOLO-format bounding boxes, consisting of normalized coordinates for the center, width, and height of each object of interest, along with the corresponding class identifier. Standard data augmentation techniques were applied, including random rotations up to 15 degrees, brightness and contrast variations up to 20%, Gaussian noise with a standard deviation of 0.01, and horizontal flipping with a probability of 50%. These techniques are established methods for improving generalization in computer vision models [18].

The selected detector was YOLOv8, which uses multi-scale feature extraction and an anchor-free detection head with separate localization and classification outputs [19]. YOLOv8 was chosen for its favorable balance between detection performance and computational efficiency, particularly for offboard inference in robotic perception applications.

Fine-tuning began with a YOLOv8 model pre-trained on the COCO dataset, which provides generalized visual knowledge for 80 common object classes. Training parameters were configured with an initial learning rate of 0.001, cosine annealing decay over 105 epochs, a batch size of 16 images, Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and early stopping with a patience of 10 epochs monitoring the mean Average Precision on the validation set [20]. For YOLOv8 detection, the optimization objective combines bounding-box regression loss, classification loss, and Distribution Focal Loss (DFL), weighted according to the Ultralytics implementation [19].

Training was performed on a workstation equipped with an Intel Core i9-12900K CPU, 64 GB of RAM, and an NVIDIA GeForce RTX 4070 GPU with 12 GB of VRAM. The process took approximately 2 hours to complete 105 epochs. The resulting fine-tuned model was exported in PyTorch format, with the `best.pt` weights file containing the optimized parameters for PPE detection. The final architecture consists of 72 layers and 3,000,818 trainable parameters, balancing depth and computational efficiency. Table 2 presents the mAP50 values obtained for each class, allowing performance analysis across PPE categories.

Table 2: mAP50 for each class

Class	mAP50
Hardhat	0.987
Person	0.955
Safety_Boots	0.924
Safety_Gloves	0.925
Safety_Mask	0.928
Safety_Vest	0.959

The mAP50 values in Table 2 are validation-set metrics obtained during model training and should be distinguished from the detection coverage later reported for the five synthetic proof-of-concept scenes. Confidence scores describe the model's confidence in individual accepted detections, whereas detection coverage describes the fraction of expected target instances identified in each controlled synthetic scene.

Integration with the NAO robot was implemented using the NAOqi SDK version 2.8, which provides a Python interface for communicating with the robot services via the qi messaging protocol [21]. The connection was established over WiFi using IP address 172.15.1.29 and default port

9559, creating a persistent session through the `qi.Session` class. Three main NAO services were utilized: ALVideoDevice for capturing images from the top frontal camera, ALTextToSpeech for announcing detected PPE, and ALMotion for controlling motor stiffness.

The camera was configured at 640×480 resolution in RGB color space with an acquisition rate of 30 frames per second, balancing image quality and processing latency. The camera subscription was made using `subscribeCamera` with a unique identifier, granting exclusive access to the video stream during system operation.

The main processing loop is implemented asynchronously. Image capture occurs continuously, while YOLO inference is executed at 2-second intervals to optimize computational resource usage and prevent overload on the robot's embedded system. Each captured frame is converted from the NAO image buffer to a three-dimensional NumPy array representing height, width, and color channels before being supplied to the image-processing and detection pipeline.

Inference is performed using the YOLO model's predict method, with a confidence threshold of 0.3 to filter low-certainty detections. Verbose logging is disabled. Inference results consist of a list of objects with bounding box coordinates, confidence scores, and class indices, which are processed to extract class names using model.names and stored in a persistent data structure. Figure 2 presents a block diagram illustrating the responsibilities of each module in this process.

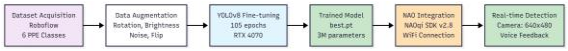


Figure 2: Block diagram of the processing pipeline.

Real-time visualization is implemented using OpenCV. Colored bounding boxes are drawn over detected objects: Person in cyan, Safety Vest in magenta, Hardhat in blue, Safety Boots in white, Safety Gloves in light cyan, and Safety Mask in black. Textual information is overlaid, including a countdown to the next detection and the total number of PPE items detected in the last cycle, providing clear visual feedback.

The ALTextToSpeech service was initialized as an interaction resource. In the evaluated implementation, quantitative assessment focused on visual detection output and inference latency; speech-based notification remains an optional interaction feature rather than an evaluated performance component.

System validation was performed using five high-resolution synthetic images generated by the Grok generative model developed by xAI, specialized in photorealistic image synthesis from detailed textual prompts. The prompts specified construction and industrial scenarios with workers wearing different PPE combinations, including multi-person scenes. This setup, where NAO observes images displayed on a monitor, constitutes a controlled proxy for field validation, enabling reproducible performance benchmarking prior to real-world deployment.

The five generated images present increasing complexity, from a single worker wearing helmet, vest, and boots to five workers using different PPE combinations, including helmets, vests, boots, gloves, and masks. Each synthetic image was submitted to the detection system, recording performance metrics, including detected classes, individual confidence scores, inference time, and bounding box coordinates.

Analysis focused on three main dimensions: confidence scores for accepted detections, processing speed quantified by inference time in seconds, and detection coverage across the controlled synthetic scenes.

It is worth noting that YOLO inference was performed offboard on a dedicated workstation, with the NAO acting as a robotic sensing and interaction node. During the reported validation, the robot remained stationary and autonomous locomotion was not evaluated. Onboard inference and autonomous navigation remain directions for future optimization.

Results were documented through OpenCV screenshots, textual logs with timestamps and detailed metrics, and tables summarizing aggregated performance by class and image.

The code utilized can be publicly found at:

[Repository URL omitted for double-blind review.]

4. MATHEMATICAL MODELING

The proposed system can be modeled as a discrete visual detection process in which an image acquired by the NAO robot is periodically processed by a trained YOLOv8 model. Let the image captured at instant k be represented by

$$\mathbf{I}_k \in \mathbb{R}^{480\times640\times3},$$

where the three channels correspond to the RGB components of the image.

The trained detector can be represented by the mapping

$$\mathcal{F}_\theta(\mathbf{I}_k) = \{(\mathbf{b}_i, c_i, p_i)\}_{i=1}^{N_k},$$

where θ represents the learned parameters of the YOLO model, $\mathbf{b}_i = (x_{1,i}, y_{1,i}, x_{2,i}, y_{2,i})$ is the bounding box of the i-th detected object, c_i is its predicted class, p_i is the associated confidence score, and N_k is the number of detections generated for frame k [4], [19].

Only detections satisfying the confidence criterion

$$p_i \geq \tau, \qquad \tau = 0.30$$

are accepted. Thus, the final detection set is defined as

$$\mathcal{D}_k = \{(\mathbf{b}_i, c_i, p_i) \mid p_i \geq 0.30\}.$$

The number of valid detections in each inference cycle is

$$N_k^{\text{det}} = |\mathcal{D}_k|.$$

Although video acquisition is continuous, inference is periodically executed with interval $T_d = 2$ s. Therefore, the nominal detection frequency is

$$f_d = \frac{1}{T_d} = 0.5 \text{ Hz}.$$

The computational latency of each YOLO inference is calculated as

$$T_{\text{inf},k} = t_k^{\text{end}} - t_k^{\text{start}},$$

where t_k^{start} and t_k^{end} correspond to the timestamps immediately before and after model execution, respectively. The complete processing flow can therefore be summarized as

$$\mathbf{I}_k \xrightarrow{\mathcal{F}_\theta} (\mathbf{b}_i, c_i, p_i) \xrightarrow{p_i \geq 0.30} \mathcal{D}_k.$$

This formulation represents the transformation of images acquired by the NAO robot into structured PPE information

containing object classes, confidence scores, and spatial locations.

5. ALGORITHM

Algorithm 1 summarizes the procedure implemented for real-time PPE detection using the NAO robot and the trained YOLOv8 model.

Algorithm 1: PPE detection using YOLOv8 and the NAO robot

REQUIRE Trained YOLO model, NAO connection, $\tau = 0.30$, $T_d = 2$ s

ENSURE Detected PPE classes, confidence scores, and bounding boxes

Connect to the NAO robot

Access camera, speech, and motion services

Configure camera at 640×480 pixels

Load the trained YOLO model

$t_{last} \leftarrow 0$

WHILE system is active

  Capture image I_k from the NAO camera

  IF image is valid

    Convert the image buffer to an OpenCV-compatible array

    $t_{current} \leftarrow$ current time

    IF $t_{current} - t_{last} \geq T_d$

      Start inference timer

      $R_k \leftarrow F_\theta(I_k)$

      Stop inference timer

      $D_k \leftarrow \emptyset$

      FOR each detection i in R_k

        IF $p_i \geq \tau$

          Store class c_i, confidence p_i, and box b_i

        ENDIF

      ENDFOR

      Generate image with bounding boxes

      Display detected PPE and inference time

      $t_{last} \leftarrow t_{current}$

    ENDIF

    Display the latest detection result

  ENDIF

  IF key q is pressed

    Stop execution

  ENDIF

ENDWHILE

Unsubscribe from the NAO camera

Close visualization windows

The implemented procedure separates continuous image acquisition from periodic object detection. Although the NAO camera provides a continuous video stream, YOLO inference is executed every 2 s to reduce computational demand. The most recent detection result remains displayed between inference cycles. This strategy allows the robot to operate as a visual sensing and interaction platform while the computationally intensive detection stage is executed on an external workstation.

6. EXPERIMENTAL OUTCOMES AND DISCUSSION

The experimental results show promising proof-of-concept behavior for automatic PPE detection in synthetic industrial scenarios, providing quantitative evidence about detection confidence, coverage, and model-level inference latency. Analysis of the first synthetic image (Figure 3) revealed successful detection of four distinct PPE classes associated with a single worker, including Safety_Vest with confidence of 0.87, Person with 0.84, Hardhat with 0.76, and Safety_Boots with 0.32, processed in an inference time of 0.02 seconds. Gloves, mask, and one safety boot were not identified, resulting in 4 detections among 8 expected target instances, corresponding to a detection coverage of 0.50.

The high confidence for vest and person indicates model robustness in detecting prominent visual features such as reflective patterns and human silhouette, while the moderate confidence for the helmet reflects challenges associated with pose and viewing angle variations. The relatively low confidence for boots, although above the 0.3 detection threshold, suggests inherent difficulty in identifying small objects located at the lower extremities, often subject to partial occlusion and low relative resolution.



Figure 3: PPE detection results for image 1.

The second image (Figure 4) presented similar results, with four detections: Safety_Vest 0.89, Person 0.88, Hardhat 0.77, and Safety_Mask 0.55, maintaining inference time of 0.02 seconds. Gloves and both boots were not identified, resulting in 4 detections among 8 expected target instances and a detection coverage of 0.50. The successful detection of Safety_Mask demonstrates the model's ability to identify facial PPE even at reduced scale, a critical consideration in environments with airborne particles or chemical vapors. The consistent confidence values for Person and Safety_Vest across the first two images indicate stability in detecting these essential classes regardless of subtle scene variations.

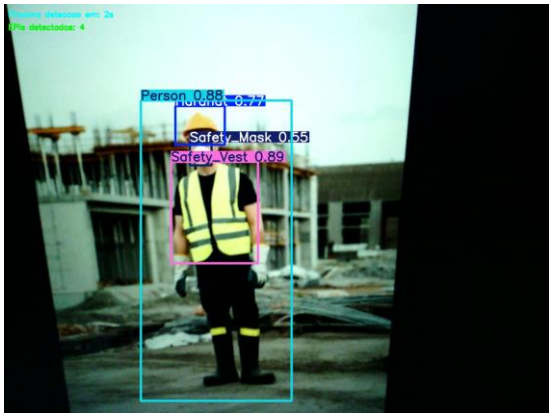


Figure 4: PPE detection results for image 2.

The third image (Figure 5) represents a qualitative increase in complexity, containing three people simultaneously, resulting in twelve total detections processed in a marginally lower inference time of 0.01 seconds. Detected instances included multiple Safety_Vest with maximum confidence of 0.88, three persons with confidence ranging from 0.82 to 0.87, two helmets with 0.77 and 0.85, two masks with 0.36 and 0.63, two gloves with 0.36 and 0.31, and one boot with 0.35. The ability to process multi-person scenes while maintaining extremely low inference time highlights the computational efficiency of YOLOv8, essential for real-time applications. However, not all expected target instances were identified, resulting in a detection coverage of 0.56. Variations in confidence among instances of the same class reflect contextual factors such as partial occlusion, relative distance to the camera, and local illumination quality.



Figure 5: PPE detection results for image 3.

The fourth image (Figure 6) maintained high complexity with three people, resulting in fifteen total detections and a slightly higher inference time of 0.04 seconds, still acceptable for real-time applications. Highlights include Safety_Vest reaching maximum confidence of 0.90, Person with 0.88, multiple Hardhats ranging from 0.31 to 0.80, and Safety_Boots with a maximum of 0.69. The slightly higher inference time is attributable to the larger number of simultaneous detections, requiring more intensive processing to eliminate redundant overlaps. The detection coverage for this inference was 0.68.



Figure 6: PPE detection results for image 4.

The fifth image (Figure 7) presented the highest complexity, with four visible people, resulting in sixteen total detections processed in 0.03 seconds. The scene demonstrated robust performance with Hardhat reaching 0.86, Safety_Vest 0.86, Person 0.86, Safety_Gloves 0.69, Safety_Mask 0.67, and Safety_Boots 0.51. The detection coverage was 0.52, as some expected target instances were not detected. The detection of multiple PPE instances in the more complex synthetic scenes indicates encouraging behavior under increased scene density, although it does not substitute for validation in real industrial environments.



Figure 7: PPE detection results for image 5.

Figure 8 presents the observed detection coverage and processing time for each inference across the five analyzed images. The graph shows an average detection coverage of 0.55, meaning that approximately 55% of the expected target instances defined for this controlled evaluation were detected. Although initial performance is promising for complex scenarios, improvements are possible. Expanding the dataset to include more examples of each PPE class across various poses, angles, and illumination conditions is an effective strategy to reduce false negatives and increase model robustness.

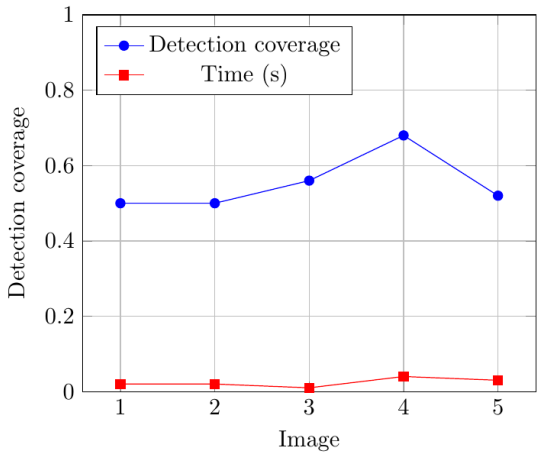


Figure 8: Observed detection coverage and model-level inference time across the five synthetic scenarios.

Aggregated analysis across the five images reveals consistent performance patterns per PPE class. Safety_Vest emerged as the most reliable class, with values consistently above 85% and a maximum of 90.05%, as expected due to distinctive visual features such as fluorescent colors and reflective geometric patterns. Person detection was equally robust, with average confidence above 85%, reflecting the benefit of pre-training on the COCO dataset, which contains thousands of diverse person examples providing transferable visual representations.

Hardhat detection was solid, with most detections above 75%, though more variable than vests, consistent with literature where helmet detection is moderately challenging due to variations in color, material, and pose [22]. Safety_Boots showed higher variability (0.32–0.69), reflecting challenges in detecting small objects at lower extremities often partially occluded by equipment or other workers. Safety_Mask and Safety_Gloves were more challenging classes, with moderate confidence values, as expected given their small size and higher visual variability compared to standardized helmets and vests.

Model-level inference times ranged from 0.01 to 0.04 seconds per image, indicating that the offboard detector itself is computationally compatible with real-time operation. However, the implemented application intentionally triggers a new inference every 2 seconds, corresponding to an effective detection-cycle frequency of 0.5 Hz. Regarding limitations, the current validation relied exclusively on AI-generated synthetic images, which may not fully capture real-world variability in lighting, occlusion, and PPE diversity. Inference was performed offboard, and the robot was not yet deployed in an active industrial facility. These constraints are acknowledged as priorities for future validation.

Direct numerical comparison with related work should be made cautiously because published studies use different datasets, targets, and evaluation protocols. For example, Wu and Zhao report a detection rate for helmet identification under their own experimental protocol [23], whereas the present proof-of-concept reports class confidence, mAP50 on the validation split, and detection coverage on five synthetic scenes. The successful integration with the NAO robotic platform demonstrates practical feasibility for deployment in real mobile monitoring contexts. The asynchronous architecture with continuous video capture and periodic inference optimizes computational resource usage while maintaining adequate system responsiveness.

The NAO platform also provides text-to-speech capabilities that can support future auditory notifications; however, speech output was not treated as an experimental performance metric in the present evaluation.

7. CONCLUSION

This work presented the full development of an automated system for detecting Personal Protective Equipment (PPE), integrating the YOLOv8 algorithm with the NAO humanoid robotic platform. The system demonstrated technical feasibility as a controlled proof-of-concept across the evaluated synthetic scenarios. The fine-tuning process on a specialized PPE dataset obtained from the Roboflow platform resulted in a model capable of detecting six classes of safety equipment with high confidence and processing speed compatible with real-time requirements.

Experimental results, evaluated on five synthetic images generated by artificial intelligence, revealed confidence rates exceeding 85% for key classes, including Safety_Vest, Person, and Hardhat. Model-level inference times ranged from 0.01 to 0.04 seconds per image, while the implemented monitoring loop executed inference every 2 seconds. These results support the computational feasibility of the offboard detection stage, while broader conclusions about field performance require real-world validation.

Integration with the NAO robot through the NAOqi framework, including video acquisition and OpenCV-based visualization, confirms the technical viability of using the humanoid platform as a robotic sensing interface for offboard PPE detection.

The detection of multiple simultaneous PPE instances in scenes containing several workers indicates that the pipeline can handle multi-person inputs, although scalability in occupied industrial facilities remains to be validated.

The main contributions of this work include a comprehensive methodology for developing a PPE detection system, from data acquisition to robotic deployment. Additionally, the study validates YOLOv8's effectiveness in the specific domain of occupational safety, demonstrates the feasibility of integrating a humanoid platform for mobile monitoring, and establishes a baseline performance protocol using synthetic images.

The obtained results position the developed system as a promising proof-of-concept for automated occupational-safety monitoring. This approach complements human inspection limitations related to fatigue, limited spatial coverage, and inability to provide uninterrupted monitoring.

Future work may focus on extensive validation in real industrial environments, comparing performance against expert human inspection. Additional directions include expanding detection capabilities to identify improper PPE usage, integrating with indoor positioning systems to generate compliance heatmaps, and developing multi-object tracking algorithms for longitudinal safety behavior analysis.

Incorporating continuous learning techniques would allow automatic model adaptation to new PPE types and environmental conditions without the need for complete offline retraining.

Extending the platform to include autonomous navigation capabilities based on SLAM would enable fully autonomous patrols, maximizing operational efficiency.

Overall, this work demonstrates the technical feasibility of combining deep learning, humanoid robotics, and synthetic test data in a controlled PPE-monitoring prototype. The results establish a baseline for future studies using real industrial imagery, formal detection metrics, and autonomous robotic inspection.

AUTHOR CONTRIBUTIONS

[Author contribution statement omitted for double-blind review.]

REFERENCES

[1] J. Hinze and J. Teizer, "Visibility-related fatalities related to construction equipment," *Safety Science*, vol. 49, no. 5, pp. 709–718, 2011, doi: 10.1016/j.ssci.2011.01.007.

[2] International Labour Organization, *Safety and health at the heart of the future of work: Building on 100 years of experience*. Geneva: International Labour Organization, 2019. Available: https://www.ilo.org/global/publications/books/WCMS_686645/lang--en/index.htm

[3] C. B. de Oliveira, W. A. Varella, A. de Carvalho Junior, and D. de M. Henrique, "Monitoramento de EPI com detecção de objetos para ambientes de trabalho," in *Anais do 15º congresso de inovação, ciência e tecnologia do IFSP (CONICT)*, 2024.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[5] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020, Available: https://arxiv.org/abs/2004.10934

[6] D. Gouaillier *et al.*, "Mechatronic design of NAO humanoid," in *Proceedings of the IEEE international conference on robotics and automation*, Kobe, Japan, 2009, pp. 769–774. doi: 10.1109/ROBOT.2009.5152516.

[7] ARDI, "PPE-HFJOC: Dataset de equipamentos de proteção individual." Roboflow Universe, 2025. Available: https://universe.roboflow.com/ardi-csjyk/ppe-hfjoc

[8] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019, doi: 10.1109/TNNLS.2018.2876865.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, San Diego, CA, USA, 2005, pp. 886–893. doi: 10.1109/CVPR.2005.177.

[10] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, HI, USA, 2017, pp. 7263–7271. doi: 10.1109/CVPR.2017.690.

[11] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018, Available: https://arxiv.org/abs/1804.02767

[12] S. Shamsuddin, H. Yussof, L. I. Ismail, S. Mohamed, F. A. Hanapiah, and N. I. Zahari, "Initial response in HRI – a case study on evaluation of child with autism spectrum disorders interacting with a humanoid robot NAO," *Procedia Engineering*, vol. 41, pp. 1448–1455, 2012, doi: 10.1016/j.proeng.2012.07.334.

[13] E. Pot, J. Monceaux, R. Gelin, and B. Maisonnier, "Choregraphe: A graphical tool for humanoid robot programming," in *IEEE international symposium on robot and human interactive communication*, Toyama, Japan, 2009, pp. 46–51. doi: 10.1109/ROMAN.2009.5326209.

[14] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *European conference on computer vision*, Zurich, Switzerland, 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.

[15] S. I. Nikolenko, *Synthetic data for deep learning*, vol. 174. in Springer optimization and its applications, vol. 174. Springer, 2021. doi: 10.1007/978-3-030-75178-4.

[16] A. Ramesh *et al.*, "Zero-shot text-to-image generation," in *Proceedings of the international conference on machine learning*, 2021, pp. 8821–8831. Available: http://proceedings.mlr.press/v139/ramesh21a.html

[17] B. Dwyer, J. Nelson, and J. Solawetz, "Roboflow: Give your software the power to see objects in images and video." 2022. Available: https://roboflow.com

[18] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019, doi: 10.1186/s40537-019-0197-0.

[19] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by ultralytics (version 8.0.0)." 2023. Available: https://github.com/ultralytics/ultralytics

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2015, Available: https://arxiv.org/abs/1412.6980

[21] SoftBank Robotics, "NAOqi developer guide." 2020. Available: http://doc.aldebaran.com/2-8/index.html

[22] Q. Fang *et al.*, "Detecting non-hardhat-use by a deep learning method from far-field surveillance videos," *Automation in Construction*, vol. 85, pp. 1–9, 2018, doi: 10.1016/j.autcon.2017.09.018.

[23] H. Wu and J. Zhao, "An intelligent vision-based approach for helmet identification for work safety," *Computers in Industry*, vol. 100, pp. 267–277, 2018, doi: 10.1016/j.compind.2018.03.037.