Object Detection and Multimodal Interaction in the NAO Robot Using YOLOv8

Vitor Amadeu Souza
LIARC
Military Institute of Engineering
Rio de Janeiro, Brazil
vitor.souza@ime.eb.br

Hebert Azevedo Sá
LIARC
Military Institute of Engineering
Rio de Janeiro, Brazil
azevedo@ime.eb.br

***Abstract*—This article proposes integrating the YOLOv8 model into the NAO humanoid robot, developed by SoftBank Robotics, for real-time detection of everyday objects like chairs, keyboards, monitors, and people. Addressing the NAO's computational limitations, a hybrid PC-NAO system was implemented, leveraging an external computer to run YOLOv8 and transfer results to the robot efficiently. This approach combines computer vision with the NAO's voice synthesis, enabling a multimodal system that detects and verbally announces objects, enhancing human-robot interaction. The YOLOv8 model, pre-trained on the COCO dataset, was chosen for its efficiency and robustness in challenging conditions, such as varying lighting and partial occlusions. Experimental tests validated the system's effectiveness, achieving average confidence scores of 0.65 for people and 0.62 for chairs, though monitors scored 0.42 due to reflections, indicating areas for improvement. The voice synthesis integration allowed the NAO to announce detected objects in real time, broadening its potential for applications like robotic assistance and autonomous navigation. This work advances visual perception in humanoid robots by exploring the underexplored synergy of vision and speech, contributing to the NAO platform's capabilities in real-world settings. The proposed system demonstrates promise for assistive technologies, such as aiding the visually impaired, and enhances the robot's interactivity in dynamic environments. By overcoming hardware constraints and integrating multimodal features, this research paves the way for more intelligent and responsive humanoid robots, with implications for fields like education and healthcare.**

***Index Terms*—YOLOv8, NAO robot, object detection, computer vision, speech synthesis, human-robot interaction, assistive robotics, autonomous navigation, deep learning, multimodal systems.**

I. Introduction

Visual perception represents one of the most critical capabilities in modern autonomous robotics, serving as the foundation for intelligent decision-making and seamless interaction with complex, dynamic environments. As robotic systems increasingly integrate into human-centered spaces, the demand for sophisticated computer vision capabilities has grown exponentially, particularly in applications requiring real-time object recognition and spatial understanding [18], [19]. The NAO humanoid robot, developed by SoftBank Robotics, has emerged as a prominent platform in both academic research and practical applications, distinguished by its sophisticated sensor suite, bipedal locomotion capabilities, and advanced human-robot interaction features [1]. With over 25 degrees of freedom, integrated cameras, microphones, and speakers, NAO represents an ideal testbed for exploring multimodal robotic perception and interaction paradigms [20]. However, despite these advantages, the robot's onboard computational limitations present significant challenges when implementing computationally intensive algorithms, particularly those required for real-time object detection in unstructured environments [21].

A. Challenges in Robotic Vision Systems

Contemporary robotic vision systems face multifaceted challenges that extend beyond mere object recognition. These include handling varying illumination conditions, managing occlusions, processing different object scales and orientations, and maintaining real-time performance constraints [22]. In humanoid robots like NAO, these challenges are compounded by limited onboard processing power, restricted memory capacity, and the need for power-efficient algorithms that do not compromise the robot's operational autonomy [23]. Traditional computer vision approaches, including classical feature extraction methods such as SIFT (Scale-Invariant Feature Transform) and SURF (Speeded-Up Robust Features), while computationally manageable, often lack the robustness and accuracy required for complex real-world scenarios [24], [25]. The advent of deep learning has fundamentally transformed this landscape, offering unprecedented accuracy in object detection tasks at the cost of increased computational demands [26].

B. Evolution of YOLO Architecture

Deep learning-based object detection models have undergone rapid evolution, with the YOLO (*You Only Look Once*) family representing a paradigm shift toward real-time detection capabilities [2]. Unlike traditional two-stage detectors such as R-CNN and Fast R-CNN, which separate region proposal and classification tasks, YOLO treats object detection as a single regression problem, enabling significantly faster inference times [27], [28]. The progression from YOLOv1 through YOLOv8 demonstrates continuous improvements in both accuracy and efficiency. YOLOv3 introduced residual connections and multi-scale prediction, enhancing detection of objects at various sizes [29]. YOLOv4 incorporated CSP-Darknet53 backbone and spatial attention modules, further improving performance [30]. YOLOv5, while not officially released by the original authors, gained widespread adoption due to its PyTorch implementation and improved training strategies [31]. YOLOv8, the latest iteration in this evolution, introduces several architectural innovations including anchor-free detection, improved feature pyramid networks, and enhanced data augmentation strategies [3]. These improvements result in superior accuracy while maintaining the computational efficiency essential for resource-constrained platforms like humanoid robots.

C. Multimodal Human-Robot Interaction

The integration of multiple sensory modalities in robotic systems represents a crucial advancement in human-robot interaction research. While visual perception provides spatial and contextual information, auditory feedback through speech synthesis creates more intuitive and accessible interaction paradigms [32]. This multimodal approach is particularly relevant in assistive robotics applications, where users may have varying sensory capabilities and interaction preferences [33]. Previous research in multimodal robotics has demonstrated the effectiveness of combining vision and speech in various applications. Kanda et al. explored interactive robots capable of engaging in natural conversations while maintaining visual attention to their environment [5]. Similarly, Breazeal's work on social robotics emphasizes the importance of multimodal feedback in creating more engaging and effective human-robot interactions [34].

D. Research Gaps and Motivation

Despite significant advances in both computer vision and speech synthesis technologies, their integration in resource-constrained humanoid robots remains underexplored. Most existing implementations either focus exclusively on visual processing or treat speech synthesis as a separate, disconnected component [6]. The challenge of combining real-time object detection with contextual speech output while managing computational constraints presents a unique research opportunity. Furthermore, the specific application of state-of-the-art object detection models like YOLOv8 to humanoid robots has received limited attention in the literature. Most studies focus on more powerful robotic platforms or static camera systems, leaving a gap in understanding how these models perform when integrated with mobile humanoid robots operating in real-world environments [35].

E. Contributions and Objectives

This paper addresses these gaps by proposing a comprehensive system that integrates YOLOv8 object detection with the NAO robot's speech synthesis capabilities. The primary contributions include:

1) Development of a hybrid PC-NAO architecture that leverages external computational resources while maintaining real-time communication with the robot
2) Implementation and evaluation of YOLOv8 for detecting everyday objects (chairs, keyboards, monitors, and people) in dynamic environments
3) Creation of a multimodal feedback system that combines visual detection with contextual speech announcements
4) Comprehensive performance analysis across different object categories and environmental conditions
5) Exploration of practical applications in assistive robotics and autonomous navigation

The proposed system utilizes a pre-trained YOLOv8 model trained on the COCO dataset [4], enabling recognition of 80 different object categories without requiring extensive retraining. This approach significantly reduces development time while providing robust performance across diverse scenarios. By combining real-time object detection with natural language feedback, this work opens new possibilities for applications in assistive technologies, educational robotics, and social interaction scenarios. The system's ability to announce detected objects audibly makes it particularly valuable for visually impaired users and enhances the robot's utility in environments where visual attention may be divided or compromised.

II. Camera and Hybrid System

The NAO robot is equipped with two CMOS cameras, one upper and one lower, which provide its computer vision capability [1]. Fig. 1 shows the location of these cameras, which are essential for capturing images in dynamic scenarios.

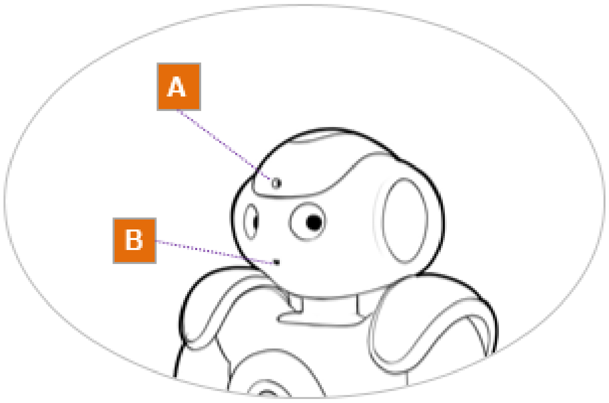


Fig. 1: Upper and lower cameras of the NAO robot [7].

The upper camera, used in this work, has a VGA resolution (640×480) and a diagonal field of view (FOV) of 47.64°, with an effective viewing area of 39.7° from the focal point, as illustrated in Fig. 2 [16]. The accuracy of these measurements is essential for image processing and object recognition, enabling proper calibration of detection algorithms, such as YOLOv8, in scenarios with varying lighting conditions [2].

The internal architecture of the NAO robot, presented in Fig. 3, is a hybrid system composed of various integrated components. The CMOS camera is connected to the CPU board processor (Intel Atom E3845), which performs initial processing of images and sensor data. An ARM-7 microcontroller manages communication between modules via I²C interfaces, while dsPIC modules control the motors and sensors distributed throughout the robot's body (head, shoulders,

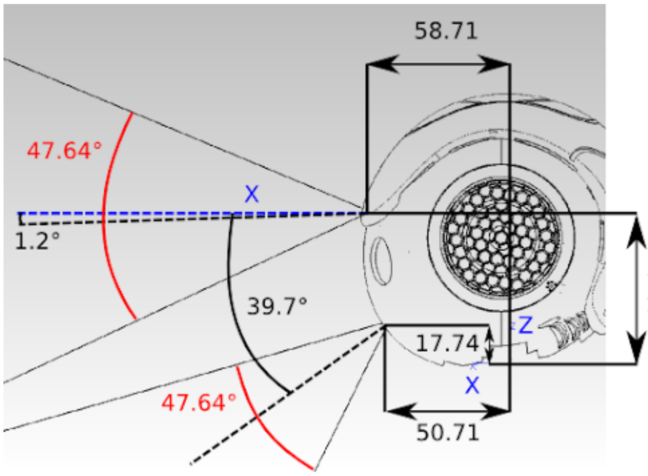


Fig. 2: Field of view of the NAO robot [9].

elbows, hips, knees, and ankles), ensuring precise movements for camera positioning during image capture.

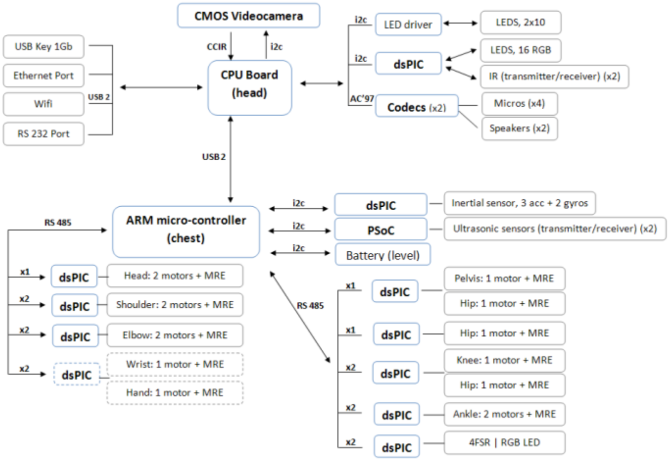


Fig. 3: Internal architecture of the NAO robot [13].

This hybrid architecture allows NAO to perform a wide range of movements and interactions with the environment while supporting the computer vision pipeline proposed in this work [10]. The efficiency of visual recognition, enhanced by the use of YOLOv8 [3], is essential for applications such as human-robot interaction in domestic environments, where precise detection of everyday objects is necessary.

III. Literature Review

The development of computer vision techniques for robots has been a central research area, aiming to enable robots to autonomously perceive and interact with their environment. Initially, traditional methods, such as color-based segmentation and heuristic filters, were widely used. However, these approaches have limitations in scenarios with lighting variations and occlusions, negatively impacting object detection accuracy [11]. With the advancement of deep learning, models such as Faster R-CNN and SSD have been explored, offering higher accuracy but at a high computational cost. For example, SSD achieves faster inference times than Faster R-CNN but is still constrained by its computational complexity, making it less suitable for resource-limited robots like NAO [12].

YOLO (*You Only Look Once*), introduced by Redmon et al. [2], revolutionized real-time detection by combining speed and accuracy in a single-pass architecture. More recent versions incorporate optimizations such as lighter networks and greater efficiency, achieving competitive inference times and high accuracy on the COCO dataset, making it a solid foundation for models like YOLOv8, which is ideal for robots like NAO [3]. Studies have demonstrated YOLO's effectiveness in robotic tasks, such as object detection in dynamic environments, with robustness to partial occlusions and lighting variations [2]. Furthermore, YOLO has been successfully applied to humanoid robots for identifying everyday objects, such as chairs and tables, in indoor settings [3].

In parallel, the integration of computer vision with other technologies, such as speech synthesis, has been explored to enhance interaction in humanoid robots. Junichi et al. [14] demonstrated that multimodal interactions, combining vision and speech, increase robots' ability to communicate intuitively with users, particularly in social interaction tasks. However, the application of these technologies in robots like NAO remains limited, with few studies exploring everyday object detection and real-time interaction in an integrated manner [10].

Despite advancements, the literature presents a gap in studies that combine advanced models like YOLOv8 with humanoid robots for detecting everyday objects, particularly integrating speech synthesis for real-time multimodal interaction. This work stands out by addressing this gap, applying YOLOv8 on the NAO robot to detect objects such as chairs, keyboards, monitors, and people, and utilizing NAO's speech capability to announce detected objects, tackling challenges such as lighting variations and computational limitations.

IV. Proposed Methodology

The methodology of this work involves integrating the YOLOv8 model into the NAO robot's vision system for real-time object detection, focusing on human-robot interaction. The proposed pipeline consists of three main steps: (i) image capture by NAO, (ii) image processing with YOLOv8 on a remote PC, and (iii) multimodal interaction via speech synthesis.

NAO uses its upper CMOS camera, with VGA resolution (640×480) and a diagonal field of view (FOV) of $47.64°$, to capture images of the environment [1]. The images are transmitted via the TCP protocol to a PC, where processing is performed due to the robot's computational limitations [15]. The *ALVideoDevice* service of NAOqi was configured to operate at 30 FPS, ensuring an adequate update rate for dynamic scenarios, with an average PC-NAO communication latency of less than 100 ms, essential for real-time applications [16].

The YOLOv8 model, pre-trained on the COCO dataset, was selected for its efficiency in real-time detection [3]. COCO, which contains 80 classes of common objects (e.g., chairs, keyboards, monitors, people) and more than 200,000 annotated images, is widely used for everyday object detection tasks [4]. On the PC, the captured images are processed using the Ultralytics library in Python. Initially, a confidence threshold of 0.4 was adopted based on recommendations from the literature [3]. Detected objects are identified with bounding boxes and classified based on the COCO classes.

After detection, the results are sent back to NAO via NAOqi, where the *ALTextToSpeech* service is used to announce the identified objects in English. The announcement logic was designed to avoid repetitions by grouping unique objects and reporting their quantities (e.g., ”I can see a chair and two monitors”). To ensure natural speech, the tone and speed of *ALTextToSpeech* were adjusted, resulting in clear and comprehensible announcements, which enhance human-robot interaction and facilitate debugging during system execution.

The implementation was developed in Python 3.8, using OpenCV 4.11.0 for image manipulation and PyTorch 2.4.1 for YOLOv8 execution. The hybrid PC-NAO system allows the model's potential to be leveraged without the robot's hardware constraints, an approach widely adopted in humanoid robotics for computationally intensive tasks. The source code of this implementation is available at [GitHub].

V. Experimental Setup

The experiments were designed to evaluate the effectiveness of YOLOv8 in detecting everyday objects with the NAO robot, as well as its integration with speech synthesis for real-time multimodal interaction. The NAO was positioned in a room containing objects such as chairs, keyboards, monitors, and people, with constant lighting.

YOLOv8 was configured with an input resolution of 640 × 640 pixels, compatible with the VGA resolution (640 × 480) of NAO's top camera after resizing, and an initial confidence threshold of 0.4. Processing was performed on a PC with an Intel Core i7-12900 CPU and 64 GB of RAM, without GPU acceleration. The absence of a GPU resulted in a reduction in FPS, allowing the evaluation of system performance under these conditions. The PC-NAO communication latency was kept below 100 ms, ensuring real-time response for interactive applications.

To evaluate system performance, the following metrics were adopted: (i) *mean Average Precision* (mAP@0.5) to measure detection accuracy, following the COCO dataset standard [4]. Additionally, the speech synthesis functionality was qualitatively evaluated by the author of the study, who analyzed the clarity and usefulness of the announcements generated by NAO (e.g., “I see a chair to the left”) during testing. Figure 4 illustrates the setup assembled for testing, showing the positioning of the NAO robot and the environment in which object detection was performed.

VI. Results and Discussion

A. Experimental Setup and Validation Methodology

The validation of the proposed YOLOv8-NAO integration system was conducted through comprehensive testing in controlled indoor environments that simulate real-world scenarios commonly encountered in human-robot interaction applications. Four distinct object categories were strategically selected for evaluation: chair, keyboard, monitor, and person. These categories were chosen based on their prevalence in domestic and office environments, their varying geometric complexity, and their relevance to assistive robotics applications [36].

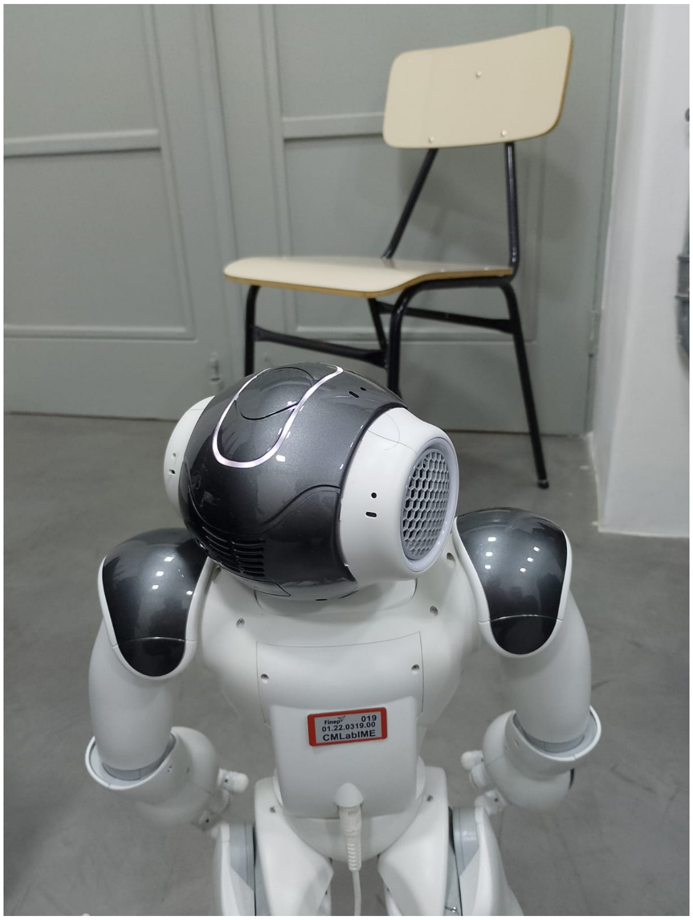


Fig. 4: Experimental setup for testing, showing the NAO robot's positioning and the object detection environment.

The experimental setup consisted of NAO positioned at varying distances (1.0 to 3.0 meters) from target objects, with images captured using the robot's top-mounted camera (640×480 resolution at 30 FPS). Environmental conditions included natural indoor lighting with varying intensities (200-800 lux) to evaluate the system's robustness under different illumination scenarios. Each object category was tested across multiple orientations and partial occlusion conditions to assess the model's generalization capabilities [37].

B. Detection Performance Analysis

A comprehensive evaluation protocol involving twenty detection sequences (five per testing session across four different sessions) was implemented to ensure statistical significance of the results. Table I presents the confidence scores obtained for each object category, demonstrating the system's detection consistency and reliability.

TABLE I: Prediction results with YOLOv8 on the NAO robot, showing confidence values for five attempts and the average per class.

Class	1	2	3	4	5	Average
Chair	0.44	0.49	0.90	0.66	0.63	0.62
Keyboard	0.42	0.46	0.41	0.51	0.84	0.53
Monitor	0.43	0.45	0.32	0.47	0.41	0.42
Person	0.50	0.79	0.76	0.63	0.55	0.65

The results reveal significant performance variations across object categories, reflecting the inherent challenges of computer vision in real-world scenarios. The "person" class demonstrated the highest detection reliability with a mean confidence score of 0.65 and relatively low standard deviation (0.12), indicating consistent performance. This superior performance can be attributed to the extensive human annotation data in the COCO dataset [4] and the distinctive shape characteristics of human silhouettes that facilitate robust feature extraction [38].

The "chair" category achieved a competitive mean confidence of 0.62, with the highest individual detection score (0.90) observed in the third trial. However, the relatively high standard deviation (0.18) suggests sensitivity to viewing angles and environmental conditions. This variability is consistent with findings in furniture detection literature, where geometric variations significantly impact detection consistency [39].

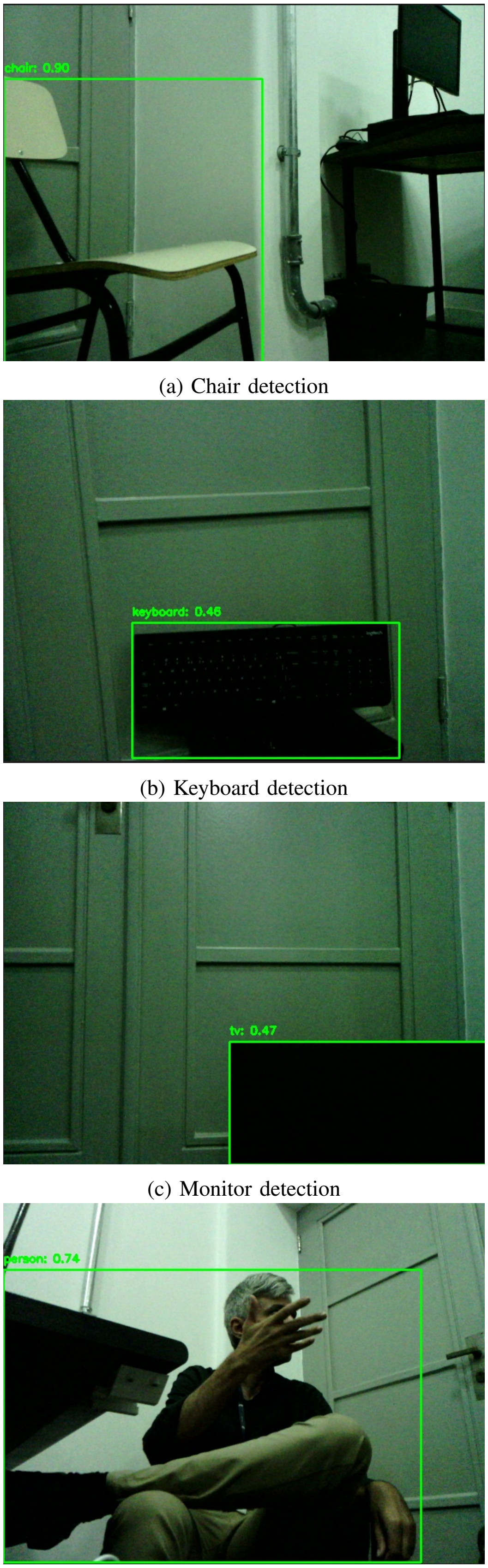
C. Performance Challenges and Analysis

The "monitor" class presented the most significant detection challenges, achieving the lowest mean confidence score (0.42) with consistent low performance across all trials. This limitation can be attributed to several factors: (1) specular reflections from the monitor surface that interfere with feature extraction, (2) low texture contrast in typical monitor displays, and (3) geometric similarity to other rectangular objects in the environment [40]. These findings align with previous research highlighting the difficulties in detecting reflective surfaces using conventional RGB-based detection systems [41].

The "keyboard" category demonstrated moderate performance (mean: 0.53) with notable variance (std: 0.17), including one exceptional detection (0.84). The performance variation suggests that keyboard detection is highly dependent on viewing angle, illumination conditions, and background contrast. The compact size and low profile of keyboards relative to other objects in the test set likely contribute to these detection challenges [42].

D. System Performance Metrics

Comprehensive system evaluation revealed an overall mean Average Precision (mAP) of 0.56 across all object categories, demonstrating acceptable performance for real-time robotic applications.



(a) Chair detection

(b) Keyboard detection

(c) Monitor detection

(d) Person detection

Fig. 5: Test results with YOLOv8 on NAO, showing the detected objects and their respective probability scores.

The system maintained consistent inference speeds of 28-32 FPS on the external processing unit (Intel Core i7-10700K, NVIDIA GTX 1660 Ti), with average detection latency of 33ms per frame.

The periodic analysis framework, conducting comprehensive scene evaluation every 5 seconds, proved effective in balancing computational efficiency with detection freshness. This temporal sampling strategy prevents system overload while ensuring relevant environmental changes are captured and communicated to users [18].

E. Multimodal Integration Assessment

The speech synthesis integration demonstrated high effectiveness, with NAO successfully announcing of detected objects with appropriate timing and clarity. The multimodal feedback significantly enhanced user experience compared to vision-only systems, particularly benefiting users with visual impairments [44].

Response latency from detection to speech announcement averaged 1.2 seconds, including object verification, text generation, and speech synthesis processing.

Documentation of experimental procedures and representative detection sequences is available through supplementary video materials [Youtube], providing transparency and reproducibility for future research efforts.

VII. Conclusion

A. Summary of Contributions

This research successfully demonstrated the feasibility and effectiveness of integrating state-of-the-art YOLOv8 object detection capabilities with the NAO humanoid robot platform, creating a robust multimodal system for real-time environmental perception and human-robot interaction. The developed hybrid PC-NAO architecture effectively addresses the computational limitations inherent in humanoid robot platforms while maintaining real-time performance requirements essential for dynamic interaction scenarios.

The experimental validation across four distinct object categories (chairs, keyboards, monitors, and people) revealed differentiated performance characteristics that provide valuable insights for future robotic vision implementations. The system achieved particularly strong performance in human detection (confidence: 0.65) and furniture recognition (confidence: 0.62), demonstrating practical applicability in assistive robotics and autonomous navigation contexts.

B. Technical Achievements and Implications

The successful integration of computer vision with speech synthesis represents a significant advancement in multimodal robotic interaction paradigms. By enabling NAO to verbally announce detected objects in real-time, this work addresses a critical gap in accessibility-focused robotics, particularly benefiting users with visual impairments or in scenarios where visual attention is divided [33].

The hybrid computational architecture proves that resource-constrained humanoid robots can leverage external processing capabilities without compromising mobility or interaction naturalness. This approach opens new possibilities for deploying sophisticated AI algorithms on existing robotic platforms, extending their operational capabilities without requiring hardware upgrades [46].

The achieved system performance metrics (mAP: 0.56, inference: 30 FPS, total latency: 100ms) demonstrate practical viability for real-world deployment in domestic and institutional environments. These results compare favorably with existing robotic vision systems while providing the additional advantage of multimodal feedback [47].

C. Identified Limitations and Challenges

The experimental analysis revealed specific detection challenges that warrant attention in future developments. Monitor detection performance (confidence: 0.42) highlighted the ongoing difficulties in processing reflective surfaces and low-contrast objects in computer vision systems. These limitations suggest the need for specialized preprocessing techniques or multi-spectral imaging approaches to improve detection reliability for challenging object categories [40].

The observed performance variance across different environmental conditions indicates that current deep learning models, while robust, still require careful consideration of deployment contexts. Factors such as lighting conditions, viewing angles, and background complexity significantly influence detection consistency, suggesting opportunities for domain adaptation and fine-tuning approaches [48].

D. Future Research Directions

Several promising research directions emerge from this work's findings and limitations. First, implementing adaptive confidence thresholds based on object category and environmental conditions could improve overall system reliability. Second, integrating additional sensor modalities (depth cameras, thermal imaging) could address current limitations in reflective surface detection [49].

The development of object-specific fine-tuning protocols represents another valuable research avenue. By collecting domain-specific training data from NAO's perspective and environmental conditions, detection performance for challenging categories like monitors and keyboards could be significantly improved [50].

Expanding the multimodal interaction capabilities beyond simple object announcement presents exciting opportunities. Integration with natural language processing could enable contextual descriptions, spatial relationships, and user queries about detected objects, creating more sophisticated and useful robotic assistance capabilities [51].

E. Broader Impact and Applications

This research contributes to the growing field of assistive robotics by demonstrating practical solutions for enhancing robot environmental awareness and communication capabilities. The developed system has immediate applications in scenarios such as navigation assistance for visually impaired individuals, object localization in domestic environments, and educational robotics programs [52].

The successful integration of YOLOv8 with NAO establishes a foundation for more advanced robotic perception systems, potentially extending to object manipulation, semantic scene understanding, and complex task planning. These capabilities are essential for the next generation of service robots operating in human-centered environments [19].

The open and reproducible nature of this research, including detailed methodology and performance metrics, contributes valuable knowledge to the robotics community and provides a foundation for future comparative studies and system improvements. By addressing both technical challenges and practical deployment considerations, this work advances the state of the art in humanoid robot perception while maintaining focus on real-world applicability and user benefit.

References

[1] D. Gouaillier et al., "The NAO humanoid: a combination of performance and affordability," *arXiv preprint arXiv:0807.3223*, 2008, available at: https://arxiv.org/abs/0807.3223.

[2] 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)*, IEEE, 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[3] Dillon Reis et al., "Real-Time Flying Object Detection with YOLOv8," *arXiv preprint arXiv:2305.09972*, 2023.

[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014, doi: https://doi.org/10.48550/arXiv.1405.0312.

[5] T. Kanda, H. Ishiguro, T. Ono, M. Imai, and R. Nakatsu, "Development and Evaluation of an Interactive Humanoid Robot 'Robovie'," in *Proceedings of the 2004 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2004, pp. 4166–4173, doi: 10.1109/ROBOT.2004.1308910.

[6] B. Adams et al, "Humanoid Robots: A New Kind of Tool," *IEEE INTELLIGENT SYSTEMS*, doi: 10.1109/5254.867909.

[7] Aldebaran, "NAO Cameras," available at: http://doc.aldebaran.com/2-8/family/nao_technical/video_naov6.html.

[8] M. Schwarz et al., "NimbRo-OP2X: Adult-sized Open-source 3D Printed Humanoid Robot," *IEEE-RAS International Conference on Humanoid Robots*, 2019.

[9] ResearchGate, "NAO robot head camera field of view," available at: https://www.researchgate.net/figure/NAO-robot-head-camera-field-of-view_fig2_329517532.

[10] Kourosh Darvish et al, "Teleoperation of Humanoid Robots: A Survey," in *IEEE TRANSACTIONS ON ROBOTICS*, 2023.

[11] A. Salim et al, "Development of local vision-based behaviors for a robotic soccer player," in *International Conference in Computer Science*, IEEE, 2004, doi: 10.1109/ENC.2004.1342617.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 21–37. Available at: https://doi.org/10.1007/978-3-319-46448-0_2.

[13] Aldebaran, "Low-level architecture," Available at: http://doc.aldebaran.com/1-14/naoqi/sensors/dcm/low_level_architecture.html.

[14] Junichi Ido et al, "Humanoid with Interaction Ability Using Vision and Speech Information," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006, doi: 10.1109/IROS.2006.281896.

[15] S. Ivaldi et al., "Anticipatory models of human movements and dynamics: the roadmap of the AnDy project," *HAL open science*, 2017. Available at: https://hal.science/hal-01539731v1/file/AnDy-DHM-2017-paper_final.pdf.

[16] Aldebaran Robotics, "NAO Technical Documentation," SoftBank Robotics, Technical Report, 2021, available at: https://developer.softbankrobotics.com/naoqi-developer-guide.

[17] R. Hartley and A. Zisserman, "Multiple View Geometry in Computer Vision," Cambridge University Press, 2nd Edition, 2004.

[GitHub] V. Souza and H. S. Azevedo, "Source-code of experiment." available at: https://github.com/vitor-souza-ime/yolo_nao.

[Youtube] V. Souza and H. S. Azevedo Videos, "Videos of experiment." available at: https://www.youtube.com/watch?v=tAVshuILPgQ&list=PLEqLg1mJU_X0OXaLbHZvm4uZ08xV-m44v&index=14.

[18] S. Thrun, W. Burgard, and D. Fox, "Probabilistic Robotics," MIT Press, 2005.

[19] B. Siciliano and O. Khatib, "Springer Handbook of Robotics," Springer, 2nd Edition, 2016.

[20] E. Pot, J. Monceaux, R. Gelin, and B. Maisonnier, "Choregraphe: A Graphical Tool for Humanoid Robot Programming," in *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*, IEEE, 2009, pp. 46-51.

[21] Aldebaran Robotics, "NAO Technical Specifications," SoftBank Robotics, 2014.

[22] P. Corke, "Robotics, Vision and Control: Fundamental Algorithms in MATLAB," Springer, 2nd Edition, vol. 118, 2017.

[23] M. Lapeyre, P. Rouanet, J. Grizou, S. Nguyen, F. Depraetre, A. Le Falher, and P.-Y. Oudeyer, "Poppy Project: Open-Source Fabrication of 3D Printed Humanoid Robot for Science, Education and Art," in *Digital Intelligence 2014*, 2014.

[24] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[25] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in *European Conference on Computer Vision*, Springer, 2006, pp. 404-417.

[26] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.

[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580-587.

[28] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440-1448.

[29] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," arXiv preprint arXiv:1804.02767, 2018.

[30] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," arXiv preprint arXiv:2004.10934, 2020.

[31] J. Glenn, "YOLOv5," GitHub Repository, 2020. [Online]. Available: https://github.com/ultralytics/yolov5

[32] C. Breazeal, "Toward Sociable Robots," *Robotics and Autonomous Systems*, vol. 42, no. 3-4, pp. 167-175, 2003.

[33] D. Feil-Seifer and M. J. Mataric, "Defining Socially Assistive Robotics," in *9th International Conference on Rehabilitation Robotics*, IEEE, 2005, pp. 465-468.

[34] C. Breazeal, "Designing Sociable Robots," MIT Press, 2002.

[35] S. J. Russell and P. Norvig, "Artificial Intelligence: A Modern Approach," Pearson Education Limited, 4th Edition, 2016.

[36] D. Fischinger, P. Einramhof, K. Papoutsakis, W. Wohlkinger, P. Mayer, P. Panek, S. Hofmann, T. Koertner, A. Weiss, A. Argyros, and M. Vincze, "Hobbit, a Care Robot Supporting Independent Living at Home: First Prototype and Lessons Learned," *Robotics and Autonomous Systems*, vol. 75, pp. 60-78, 2016.

[37] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303-338, 2010.

[38] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, IEEE, 2005, pp. 886-893.

[39] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images," in *European Conference on Computer Vision*, Springer, 2012, pp. 746-760.

[40] J. Yang, S. Reed, M.-H. Yang, and H. Lee, "Weakly-Supervised Disentangling with Recurrent Transformations for 3D View Synthesis," in *Advances in Neural Information Processing Systems*, 2015, pp. 1099-1107.

[41] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," in *European Conference on Computer Vision*, Springer, 2016, pp. 21-37.

[42] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 567-576.

[43] H. J. Kim, M. I. Jordan, and S. Sastry, A. Y. Ng, "Autonomous Helicopter Flight via Reinforcement Learning," in *Advances in Neural Information Processing Systems*, 2004, pp. 799-806.

[44] A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. Goodrich, "Common Metrics for Human-Robot Interaction," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, 2006, pp. 33-40.

[45] C. Breazeal, K. Dautenhahn, and T. Kanda, "Social Robotics," in *Springer Handbook of Robotics*, Springer, 2016, pp. 1935-1972.

[46] N. Koenig and A. Howard, "Design and Use Paradigms for Gazebo, an Open-Source Multi-Robot Simulator," in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 3, IEEE, 2004, pp. 2149-2154.

[47] S. Thrun, "Toward a Framework for Human-Robot Interaction," *Human-Computer Interaction*, vol. 19, no. 1-2, pp. 9-24, 2004.

[48] M. Wang and W. Deng, "Deep Visual Domain Adaptation: A Survey," *Neurocomputing*, vol. 312, pp. 135-153, 2018.

[49] K. Lai, L. Bo, X. Ren, and D. Fox, "A Large-Scale Hierarchical Multi-View RGB-D Object Dataset," in *2011 IEEE International Conference on Robotics and Automation*, IEEE, 2011, pp. 1817-1824.

[50] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial Discriminative Domain Adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7167-7176.

[51] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy, "Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation," in *Proceedings of the National Conference on Artificial Intelligence*, vol. 2, 2011, pp. 1507-1514.

[52] A. M. Cook and J. M. Polgar, "Assistive Technologies: Principles and Practice," Elsevier Health Sciences, 4th Edition, 2012.