Design and Development of a Vision System for Autonomous Item Retrieval in Warehouse Robotics

Author/s: Jordan Thuo[1] (N11046252)

Queensland University of Technology

EGB320: Mechatronics Design 2

Team 9

Jordan Thuo, Edmond Chan, Shreel Patel, Sonal Chand

Executive Summary

The rapid growth of e-commerce and just-in-time inventory systems has increased demands on warehouse operations, driving the need for faster, more accurate, and scalable solutions. Traditional manual item-picking systems face high error rates and slow processing, limiting logistics efficiency. Autonomous robots are emerging as a solution, capable of automating labor-intensive tasks; however, creating robots that can reliably navigate, identify items, and execute precise operations in complex warehouse environments presents significant challenges, especially with variable lighting, limited space, and dynamic obstacles.

This report introduces the TRL 3 PicknPackRobotics prototype, an autonomous mobile robot developed for warehouse automation. Central to its operation is the vision subsystem, which leverages advanced computer vision techniques such as HSV-based color segmentation, homography-based spatial estimation, and a multithreaded processing pipeline, to enable precise item retrieval and navigation. The system is configured for a mock warehouse environment as a proof-of-concept for potential investors, with the goal of scaling to full-scale warehouse operations.

The vision system demonstrates robust object detection under varying lighting, accurate spatial mapping for essential surfaces, and efficient real-time processing with high frame rates. Testing revealed an object detection accuracy of 98.3% under standard lighting and spatial accuracy within ± 3 cm for distance and ± 1.5 degrees for bearing at close ranges. While low lighting, motion blur, and occlusions presented challenges, mitigations such as staged verification techniques improved reliability. Future iterations may benefit from integrating machine learning or additional sensors for scalability.

In conclusion, the PicknPackRobotics prototype shows strong promise for improving warehouse efficiency, with its vision subsystem providing accuracy, scalability, and adaptability. This system establishes a solid foundation for future innovations that could significantly enhance the speed and efficiency of warehouse operations.

1.0 Introduction

The rise of e-commerce and just-in-time inventory systems has significantly increased the demands on warehouse operations in Australia, particularly in terms of speed, accuracy, and scalability (Starr et al., 2024). Traditional manual item-picking processes struggle to meet these demands due to high error rates, time consumption, and the repetitive nature of the tasks, resulting in inefficiencies that undermine the responsiveness required by modern logistics (Sheryl, 2023). Autonomous robots have emerged as a promising solution to automate these labour-intensive operations. A key enabler of this shift is computer vision technology, which allows robots to interpret visual data, identify target items, and navigate complex environments with precision (Pankratova, 2024). However, building robots capable of consistently achieving these tasks poses several technical challenges, especially in real-world warehouse environments where space, lighting, and obstacles are highly variable.

This report details the development of a Technology Readiness Level (TRL) 3 prototype mobile robotic platform designed by PicknPackRobotics. The system autonomously retrieves items from warehouse shelves and transports them to a designated packing station in a controlled, mock warehouse environment. The primary goal is to demonstrate the capabilities of the robot as a proof-of-concept to potential investors, with the eventual aim of scaling the system for fully automated, large-scale warehouse operations. This report focuses on the prototype's computer vision subsystem targeting key functionalities, ensuring that the system can be adapted for real-world implementation.

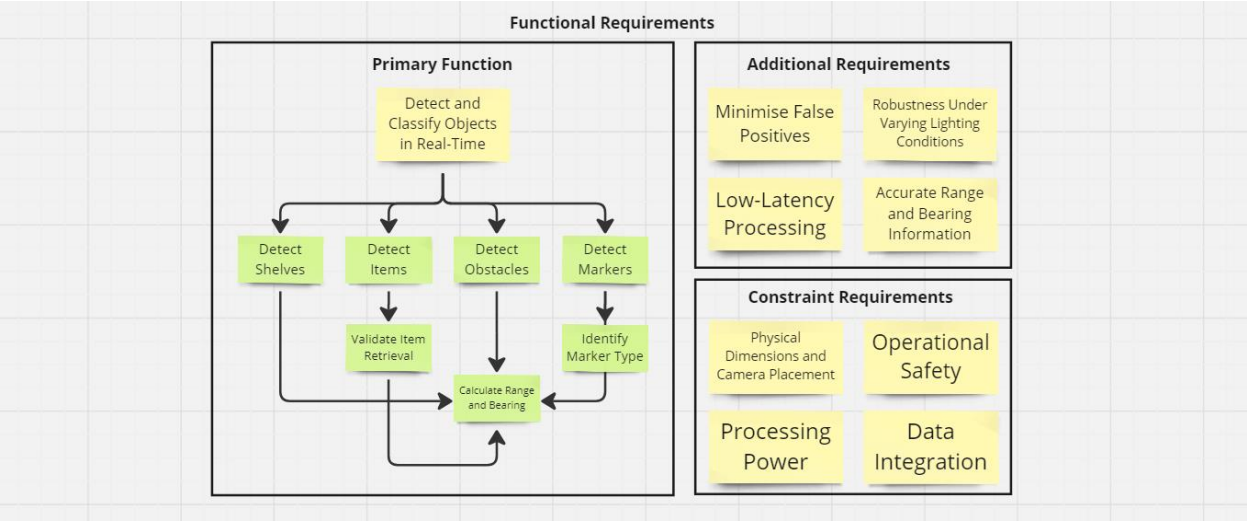

Figure 1: Key Functional Requirements for the Vision Subsystem

At the core of the robot's functionality is its vision system, which plays a critical role in detecting, classifying, and localising objects within the warehouse. By providing accurate, real-time spatial data on both range and bearing, the vision subsystem enables the robot to autonomously navigate and execute precise item retrieval tasks. Operating within the constraints of a mock warehouse environment, the system must contend with real-world challenges such as operational safety, varying lighting conditions, and processing power. To address these complexities, the vision system is designed for robustness,

modularity, and scalability, ensuring adaptability for future applications. As a result, several key functional and constraint requirements, as shown in Figure 1, were considered in its development:

1.1 Lighting Variability and Visual Occlusions: In a large-scale warehouse environment, lighting is often inconsistent, and items may be partially obscured by other objects or shelving. This can severely degrade detection accuracy in traditional vision systems. To overcome this, the prototype employs HSV-based colour segmentation, which is more resilient to changes in lighting intensity and shadows than conventional RGB-based systems (Mamdouh, 2020). While HSV improves performance, its limitations include susceptibility to extreme lighting variations, which are mitigated in this system through additional shape-based filtering and secondary confirmation checks to reduce false positives and improve robustness in challenging conditions.

1.2 Accurate Range and Bearing Estimation: Reliable spatial estimation is essential for the robot's autonomous navigation and item retrieval in tight spaces. The vision system must estimate range within a tolerance of ±5 cm and bearing within ±2 degrees, ensuring the robot aligns itself correctly with items for retrieval. By using homography to calculate distances based on known geometric relationships, the system translates 2D image coordinates into real-world spatial data. Additionally, size-based distance estimation is employed for smaller objects, providing accurate spatial information in a variety of object configurations.

1.3 Real-Time Processing and Computational Efficiency: Efficient real-time processing is critical for navigating a dynamic environment and performing item-picking tasks without delays. The vision system must process camera frames at a minimum of 10 Hz to ensure continuous operation. Delays in detection or spatial estimation could lead to missed items or collisions, which would degrade the system's overall performance. The vision system's multithreaded processing architecture optimizes frame rate and minimizes latency by parallelizing key image processing tasks, ensuring computational efficiency even when dealing with complex scenes.

1.4 Modularity and Scalability: Beyond the immediate goal of demonstrating functionality in a controlled 2x2 m mock warehouse, the system is designed for scalability. Designing the system to be modular and object-oriented enables future integration of additional sensors (such as LiDAR or depth cameras) and more advanced processing algorithms (such as machine learning for object classification). This modular approach, facilitated by clearly defined interfaces and extensible architecture, allows the system to evolve without requiring a complete redesign. As a result, the vision system is adaptable to larger, more complex environments, where dynamic elements like moving workers and varying surface types will need to be accounted for.

Principally, the vision subsystem of the TRL 3 PicknPackRobotics prototype provides precise real-time spatial data necessary for navigation and item manipulation. The system's robust handling of environmental variability, scalable modular design, and real-time processing capabilities position it as a key component for future large-scale warehouse automation. By addressing critical challenges such as

lighting variability, spatial estimation, and computational efficiency, the system is designed to evolve alongside increasing demands for warehouse automation.

2.0 Literature Review

Autonomous warehouse robots rely heavily on computer vision systems to identify, locate, and retrieve items from storage shelves (Pankratova, 2024). These vision systems must integrate advanced techniques for object detection, spatial estimation, and real-time processing to ensure robust operation in complex environments. This section reviews the current state of the art in vision systems for autonomous robots, with a focus on methods for object detection, range and bearing estimation, and system scalability.

2.1 Object Detection in Autonomous Robots

In autonomous vehicles, real-time object detection is essential for navigating through dynamic environments safely, while in warehouses, it supports tasks such as item retrieval and obstacle avoidance. Object detection methods commonly used in robotics include Convolutional Neural Networks (CNNs) such as the YOLO (You Only Look Once) series. YOLO is particularly effective for real-time applications, as it processes images in a single pass, allowing multiple objects to be detected in each frame (Pilarski et al., 2023). This is a crucial feature for environments dense with objects, such as warehouse shelves or busy traffic scenarios.

However, while CNN-based detection systems like YOLO offer high accuracy, they require significant computational resources, making them less practical for resource-constrained platforms (Othman et al., 2018). In response to this limitation, simpler methods such as HSV (Hue, Saturation, Value) colour segmentation are often applied in structured warehouse settings. HSV-based segmentation provides robust performance in variable lighting by isolating objects based on colour, a technique shown to work effectively in controlled environments like warehouses (Mamdouh, 2020). This approach aligns well with lower-power embedded systems, which prioritize processing efficiency over extensive computational power, making it suitable for warehouse robots operating in stable environments with predictable lighting conditions.

2.2 Range and Bearing Estimation

Accurate spatial estimation is essential for autonomous robots to successfully navigate and manipulate objects. complex spatial estimation techniques employ stereoscopic cameras, which estimate depth by calculating the disparity between images captured by two cameras set at a calibrated distance apart (Chen, Xu, & Ma, 2023). This method provides highly accurate depth information and has been widely adopted in robotic systems requiring precise spatial measurements. However, stereo vision systems demand more processing power and complex calibration, making them less suitable for robots operating under tight resource constraints (Taha et al., 2021).

Monocular cameras are widely used in mobile robotics due to their simplicity and low cost, but they lack inherent depth perception. This limitation can be mitigated by relying on known points of interest, such

as object size and position, to estimate distance through triangulation methods, thus enabling basic 3D scene reconstruction (Taha et al., 2021). This approach allows mobile robots with objects at various depths, despite using a single camera.

2.3 Real-Time Processing and Multithreading

Real-time processing is essential in robotics, particularly in dynamic environments where objects must be detected and localized continuously as the robot moves. Recent research has proposed powerful real-time measurement systems for industrial applications, leveraging Computer Vision techniques to detect and measure objects in video streams. Systems utilizing the Canny edge detector, enhanced with morphological operations to reduce noise and refine object outlines, have proven effective in achieving real-time performance (Othman, et al., 2018). These improvements maintain the integrity of object boundaries while enhancing shape definition, enabling fast and reliable measurements of multiple objects within a single frame.

For instance, a study implementing these methods on a Raspberry Pi 3 demonstrated the Canny edge detection with morphological operations, the system could process five frames per second, achieving a balance between speed and accuracy on a low-cost embedded platform (Othman, et al., 2018). Additionally, multithreading techniques are commonly used to optimize system performance by parallelizing tasks such as image acquisition, processing, and navigation, minimizing latency and ensuring that the system can respond quickly to changes in the environment (Shammi, et al., 2018).

In conclusion, the literature and existing designs highlight the importance of choosing the right balance between simplicity and performance. For the TRL 3 Pick-n-Pack robot, the combination of HSV segmentation, monocular vision, and multithreading provides an efficient, scalable solution suitable for the mock warehouse environment. However, future expansions will likely integrate more advanced sensors and algorithms to meet the demands of larger, more complex warehouses.

Design

The TRL 3 Pick-n-Pack robot's vision system was designed using a comprehensive, modular, and object-oriented architecture that integrates seamlessly with the navigation and item-picking subsystems. The system's design leverages advanced computer vision techniques for real-time object detection and spatial estimation, providing the necessary data for autonomous navigation in a constrained warehouse environment. This section outlines the design of the vision system, its modular architecture, and how it interfaces with other robotic subsystems, showcasing its robustness and scalability for future implementations.

3.2 Hardware System Design

The vision subsystem's hardware setup consists of a Raspberry Pi 3 as the main processing unit, the Raspberry Pi Camera V2, and a 3D-printed detachable mount. The Raspberry Pi 3 was selected for its balance between computational power and low cost, the Raspberry Pi 3 serves as the control hub for the

vision subsystem, handling image capture, processing, and communication with other subsystems. Its quad-core processor and relatively high RAM make it suitable for multithreaded processing, essential for maintaining the required frame rate of greater than 10 Hz. The camera is attached to a detachable, 3D-printed mount, which underwent several design iterations to ensure optimal positioning and flexibility. The mount's modular design allows easy adjustments to the camera's angle and height, enabling detection across various shelf heights as seen in figure 2. By optimizing the mount design, the camera maintains a wide field of view (FOV), improving item detection and spatial awareness within the robot's operating environment.

Figure 2: Close-up Image of the Final Product

3.2 System Integration and Multithreaded Architecture

The vision subsystem integrates with the robot's navigation and control units, creating a cohesive system that allows for responsive, real-time operation. Communication between the vision and navigation subsystems is achieved via a bitmask that specifies relevant object types based on the robot's current state, optimizing processing efficiency. This task-specific detection strategy reduces computational load, ensuring each subsystem receives tailored spatial data without unnecessary processing delays.

The multithreaded architecture allows parallel execution of image acquisition, and object detection. By separating these processes into individual threads, the system minimizes latency and maintains consistent frame rates, even during computationally intensive tasks. This real-time adaptability is essential for continuous movement and interaction in dynamic warehouse environments.

An image acquisition thread captures frames continuously, with each image being resized and processed for efficient computation. The camera's parameters, such as analogue gain, colour gain, and exposure time, are calibrated to balance image quality and processing speed. A transform is applied to flip the image and adjust the resolution to optimize the field of view (FOV) while maintaining computational efficiency.

The OOP design of the Camera Module ensures that camera initialization, frame capture, and configuration are separated into self-contained methods, allowing the vision system to adapt to different environments or image acquisition needs with minimal code adjustments. The captured frames are stored in a buffer, where they are processed by the vision pipeline, ensuring that no frames are missed, even during high-computation phases.

3.3 Object Detection and Processing Pipeline

The object detection component of the vision subsystem leverages HSV-based color segmentation to detect items, markers, shelves, and obstacles. Unlike traditional RGB methods, HSV segmentation is more resilient to variable lighting, as it isolates object hues effectively, making it ideal for the controlled lighting in warehouse environments.

The processing pipeline consists of the following key steps:

1. **Image Acquisition:** A thread runs camera capture in a single loop in real-time and stores the frame in a buffer to be processed in a separate thread.
2. **Object Detection:** Based on the robot's operational state (e.g., navigating to a shelf, identifying an item), a bitmask filters detected objects, focusing only on task-relevant object types, which optimises computational efficiency.
3. **Preprocessing:** To reduce noise, a Gaussian blur is applied to the captured image if required by the state. The system then applies colour thresholding in the HSV colour space, segmenting objects into a bitmask image, which remains stable under different lighting conditions.
4. **Morphological Operations:** A combination of erosion and dilation operations are used to open and close the mask, enhancing the detection accuracy of object contours by removing noise and filling small gaps. An interface is used to calibrate kernel sizes for each object, ensuring no redundant preprocessing occurs.
5. **Contour Detection and Feature Analysis:** The contours are analysed for shape, size, and position, with specific techniques like circularity and convex hull analysis helping to classify shapes and improve edge detection for shelves and markers. Markers also apply a logical bitwise and with a filled wall mask to reduce the impact of false negatives.
6. **Calculate Range and Bearing:** The order file assists the vision system to calibrate the spatial calculations to update the known dimensions of an item which allow for accurate distance measurements.

This pipeline, shown in the flowchart in Figure 3, allows the system to efficiently process frames, detect objects, and relay critical spatial information to the navigation subsystem.

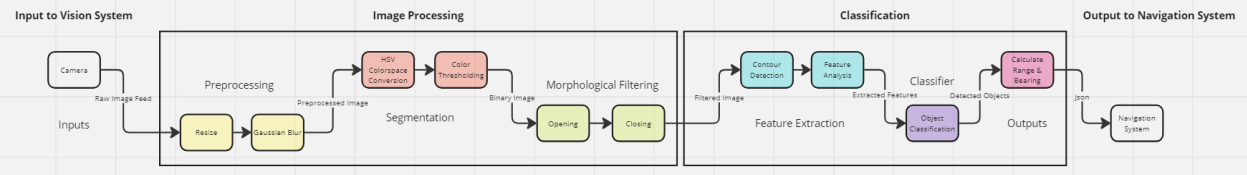

Figure 3: Image Processing Pipeline from Input to Output

3.3 Marker Detection and Wall Segmentation

The Marker Detection subsystem is crucial for accurate navigation, providing the robot with fixed reference points for localization. The system applies colour thresholding to segment walls from the scene, using HSV thresholds to isolate the white walls within the mock warehouse environment. A Convex Hull Algorithm is then applied to define the outer boundary of the wall contours, creating a Filled Wall Mask. This mask, which encapsulates the entire wall area, is then used in a Bitwise AND operation with the frame, ensuring that markers are only detected on white walls. By reducing the area of interest, the system filters out false positives, improving the reliability of marker detection.

The system further applies Shape Analysis to the detected markers, using geometric properties like circularity and aspect ratio to differentiate between square and circular markers. This is crucial for ensuring that only valid markers are recognized, minimizing false negatives. The result is an accurate and robust detection pipeline that operates with minimal computational overhead while providing essential feedback to the navigation system.

3.4 Range and Bearing Estimation

Accurate Range and Bearing Estimation is fundamental for the robot's ability to navigate toward shelves and retrieve items. The system employs two complementary methods for this purpose:

- **Homography:** A well-established technique in computer vision, homography leverages known geometric relationships between the camera and the environment to estimate distances to large, planar surfaces, such as shelves. Homography transforms the image coordinates of detected shelf corners into real-world coordinates, providing accurate range and bearing information. This method is particularly effective when working with structured environments, as it assumes a constant geometric relationship between the camera and the target.
- **Size-Based Distance Estimation:** For smaller objects or markers, the system uses size-based estimation, comparing the apparent size of an object in the frame with its known real-world dimensions to calculate its distance from the camera. This method is effective in cases where geometric relationships are less predictable, such as with small items placed on shelves.
- **Bearing Estimation:** The system calculates bearing by measuring an object's offset from the image center and converting this pixel distance into an angular value based on the camera's field of view.

  $$Bearing = (Object\ x\ Coordinate - Image\ Centre) \times Angle\ per\ pixel$$

  Where the image centre is half of the image width (e.g., 320 pixels for a 640-pixel-wide image).

3.6 Summary of Design

The theoretical basis for the vision system in autonomous robots encompasses several core areas of computer vision, such as image segmentation, edge detection, and real-time processing. HSV-based colour segmentation allows for robust object detection under variable lighting by converting RGB images into HSV colour space, where hue information remains relatively stable across lighting changes. This makes it ideal for environments where consistent lighting cannot be guaranteed, such as warehouses with overhead lighting that may create shadows and reflections.

Edge detection, enhanced by morphological operations, refines object boundaries, eliminating noise and preserving objects within a scene. OpenCV's library of vision functions, when optimized through morphological operations, proves effective for real-time applications by maintaining a clear definition of object outlines. This approach has been validated in research on embedded systems, demonstrating its ability to achieve high processing speeds without sacrificing accuracy in detecting complex object shapes.

To ensure real-time performance, the vision system is built on a multithreaded architecture that parallelizes tasks, such as image acquisition processing. By dedicating specific threads to different tasks, the system minimizes latency, maintaining a high frame rate essential for autonomous navigation and item retrieval in fast-paced environments.

Results

The vision system of the PicknPackRobotics TRL 3 prototype was thoroughly evaluated under various real-world conditions in a mock warehouse setting, examining key performance areas: object detection accuracy, spatial estimation precision, frame rate performance across operational states, and marker detection robustness. The results provide insights into both the system's strengths and areas for future improvement, establishing a foundation for scaling the system to fully automated warehouse operations.

4.1 Object Detection Accuracy

Object detection performance was tested across multiple lighting conditions to assess the robustness of the HSV-based colour segmentation approach, combined with morphological operations. A sample of 6 object types were positioned in each shelf in standard, and low light conditions, with each setup repeated 10 times to establish statistical significance.

Lighting Conditions	Object Type	Accuracy (%)
Standard Lighting	Black Marker	95%
Standard Lighting	Blue Shelf	100%
Standard Lighting	Orange Item	100%
Standard Lighting	White Wall	95%
Standard Lighting	Green Obstacle	100%
Standard Lighting	Yellow Packing Bay	100%
Poor Lighting	Black Marker	50%

Poor Lighting	Blue Shelf	100%
Poor Lighting	Orange Item	95%
Poor Lighting	White Wall	50%
Poor Lighting	Green Obstacle	80%
Poor Lighting	Yellow Packing Bay	100%

The system achieved high detection accuracy with little to no false positives and negatives in standard lighting on average 98.3%. However, accuracy decreased under inconsistent lighting conditions, where darker shadows were prevalent, reducing accuracy to 50% for markers. While the preprocessing pipeline reduced noise effectively, the system struggled with edge cases, particularly when object boundaries were not clearly defined. This trend suggests that while HSV-based segmentation is effective for vibrant coloured objects and in controlled lighting conditions, its limitations become apparent in dynamic environments. State-of-the-art deep learning methods, such as YOLOv5, tend to maintain higher accuracy and consistency across lighting conditions, indicating that integrating machine learning could improve robustness.

4.2 Spatial Estimation Precision

Spatial estimation was tested with two approaches: homography-based distance and bearing estimation for large surfaces like shelves, walls, and the packing bay, while size-based distance estimation was used for smaller objects, such as markers, items, and obstacles. These tests covered varying distances and heights to simulate realistic warehouse conditions where precise spatial awareness is crucial.

4.2.1 Homography-Based Distance and Bearing Accuracy

Homography-based estimation was applied to calculate distances and bearings for walls, the packing station, and shelves. Tests were conducted at multiple distances (0.5m, 1m, and 1.5m) to assess accuracy.

Distance (m)	Object Type	Distance Error (cm)	Bearing Error (degrees)
0.5	Packing Bay	± 0	± 0
0.5	Shelf	± 0	± 1
0.5	Wall	± 1	± 0
1	Packing Bay	± 2	± 0
1	Shelf	± 1	± 0
1	Wall	± 2	± 0
1.5	Packing Bay	± 2	± 0.5
1.5	Shelf	± 2.5	± 0.5
1.5	Wall	± 3	± 1

Homography-based estimation provided accurate results at closer ranges (± 0.3cm error average at 0.5m). However, error increased with distance, reaching ± 3cm and ± 1 degree for shelves at 1.5m. This limitation is typical of homography methods, where perspective distortion reduces accuracy at longer distances. Depth sensors could mitigate this limitation by offering consistent distance accuracy over greater ranges.

4.2.2 Size-Based Distance and Bearing Accuracy for Markers and Obstacles

Size-based distance estimation was applied to markers and obstacles across varying distances to evaluate performance at close and mid-ranges.

Object Type	Distance (m)	Distance Error (cm)	Bearing Error (degrees)
Marker	0.2	± 1.5	± 0
Marker	0.5	± 0.5	± 0
Marker	1	± 1	± 0
Obstacle	0.2	± 2	± 0
Obstacle	0.5	± 2	± 0.5
Obstacle	1	± 3	± 1.5

Size-based estimation was accurate at shorter distances but showed increased error as distance increased, reaching ± 3cm for obstacles and ± 1.5 degrees in bearing error. Stereo vision or LiDAR integration could improve the system's accuracy for small objects over longer distances.

4.2.3 Size-Based Distance and Bearing Accuracy for Items at Varying Heights

To assess distance estimation accuracy at varying heights, items were placed on three different shelf levels (Level 1, Level 2, and Level 3).

Height Level	Object Type	Distance Error (cm)	Bearing Error (degrees)
Level 1	Cube	± 2	± 0
Level 2	Bottle	± 0.5	± 0
Level 3	Ball	± 1.5	± 0
Level 1	Weetbots	± 2.5	± 0
Level 2	Mug	± 0.5	± 1
Level 3	Bowl	± 1.5	± 0

The system performed well for the middle shelf ± 0.5cm, but error increased for lower levels due to the 3D orientation of the object, affecting the accuracy of size-based distance estimation. This suggests that the system could benefit from adaptive processing for items at lower heights, further validating its utility in real-world, multi-level warehouse setups.

4.3 Frame Rate Performance by Operational State

The real-time performance of the vision system was a critical metric, particularly as it directly impacted the robot's ability to navigate and make decisions without delay. The system's multithreaded architecture successfully maintained an average frame rate of 12Hz in complex states and up to 35 Hz in simpler states, seen in figure 4, exceeding the minimum requirement of 10 Hz. This frame rate was sufficient to provide timely feedback for navigation and object detection, allowing the robot to operate smoothly.

Figure 4: Running Average Frame Rate by State a) SEARCH FOR SHELF b) SEARCH FOR ITEM c) MOVE TO ROW

Despite this success, the system encountered occasional frame drops during more computationally intensive tasks, such as detecting multiple overlapping objects or processing high-noise environments. These frame drops, though infrequent, introduced minor delays in the robot's decision-making process, potentially affecting its ability to navigate accurately in real-time. The one-frame buffer used in the image processing pipeline helped to mitigate the effects of these frame drops, but future iterations could explore more efficient processing algorithms or offloading some of the computational load to dedicated hardware, such as GPUs.

4.4 Marker Detection: Motion Blur and Occlusion

During preliminary testing, the effects of motion blur and occlusion significantly impacted the reliability of the vision system. Therefore, marker detection performance was tested under varying motion speeds and occlusion levels.

Condition	Success Rate (%)	Standard Deviation (%)	Confidence Interval (95%)
No Motion, No Occlusion	95	± 5	$\%0.5$
Fast Rotation, No Occlusion	70	± 5	$\%0.5$
No Motion, 50% Occlusion	40	± 5	$\%2.5$
Fast Rotation, 50% Occlusion	15	± 5	$\%5$

Detection success rates were high under static conditions but dropped significantly with occlusion and motion blur, indicating limitations in the bitmask approach. Future iterations could incorporate YOLO or stereo vision to address occlusion more effectively.

4.6 Final Validation and Performance in Real-World Scenarios

The vision system was tested in full-cycle trials, where the robot was tasked with autonomously completing an order consisting of six randomly placed items. Across 10 trials, the robot successfully

completed the task in 9 cases, with an average completion time of 3 minutes and 37 seconds. The single failure was caused by the system's difficulty in detecting a marker that was partially occluded by an obstacle, highlighting a key limitation in the current approach to object segmentation.

These results validate the effectiveness of the system in controlled environments, but also point to areas where further enhancements are needed. In particular, the system's sensitivity to occlusions and edge-case scenarios suggests that additional detection methodologies and improved sensor integration will be essential for scaling the design to more complex, real-world warehouse environments.

Conclusion

The development of the TRL 3 PicknPackRobotics prototype has provided a strong proof-of-concept for the role of autonomous robots in transforming warehouse operations. The project aimed to address core challenges in real-world logistics such as speed, accuracy, and scalability, by implementing a vision-driven robotic system capable of autonomously navigating and retrieving items within a structured warehouse environment. Through a modular, adaptable design focused on robustness and computational efficiency, the vision subsystem has demonstrated the potential to meet these demands effectively, though with certain limitations highlighted by real-world testing.

5.1 Summary of Achievements and Key Findings

The vision subsystem's integration of HSV-based colour segmentation and dynamic spatial estimation, and multithreading has proven highly effective in controlled environments. By leveraging these techniques, the system achieved reliable object detection, accurate range and bearing estimation, and real-time processing speeds necessary for autonomous item retrieval and navigation. Our tests showed that the system maintained high detection accuracy (98.3%) in standard lighting, with consistent spatial accuracy, achieving less than ± 3cm distance and ± 1.5 degrees bearing. Moreover, the multithreaded architecture allowed the system to maintain an optimal frame rate of 12–15 Hz in complex scenes, ensuring timely data flow and decision-making capabilities for smooth operation.

The prototype performed well in standard conditions, successfully completing order retrieval tasks in 90% of order retrieval trials. This accomplishment demonstrates that the combination of HSV-based segmentation and size-based distance estimation can be effective for item identification and spatial awareness in warehouse automation, especially in controlled lighting conditions and with stable shelving configurations. The 3D-printed, detachable mount for the camera further optimized performance by maximizing the camera's field of view and improving adaptability for different shelf heights, making it a promising solution for real-world warehouse setups.

5.2 Discussion of Limitations

While the system performed admirably in ideal conditions, testing also exposed limitations related to lighting variability, occlusions, and motion blur. The HSV-based segmentation proved highly sensitive to

lighting inconsistencies, particularly in shadows and low light scenarios. Accuracy dropped significantly to 50% in low-light conditions for certain objects, such as row markers, underscoring the need for improved lighting tolerance. Furthermore, while the homography and size-based distance estimation techniques maintained acceptable error margins at closer ranges, the accuracy diminished as distance and object complexity increased, suggesting that alternative methods, like depth sensors, may be necessary for larger-scale deployments.

The challenges posed by motion blur and occlusions were also substantial. Motion blur during quick rotations caused the vision system to miss markers, and partial occlusions affected detection reliability, especially when using bitmask filtering dependent on wall segmentation. Although staged checks were implemented to reduce the impact of motion blur, it became evident that occlusion handling requires a more advanced approach, possibly through machine learning algorithms or multi-sensor fusion, such as LiDAR integration. These limitations are critical in envisioning the next steps for scaling the system beyond the mock warehouse environment.

5.3 Future Directions and Recommendations

To extend the capabilities of the PicknPackRobotics system, several improvements are recommended based on the insights gained:

- **Enhanced Lighting Tolerance:** Incorporating deep learning models like YOLOv5 or SSD, which are less affected by lighting variability, could improve detection reliability across diverse lighting conditions. These models may be integrated alongside HSV segmentation, activating only when adverse lighting is detected.
- **Improved Spatial Estimation:** Depth sensors, such as stereo cameras or LiDAR, would help improve range accuracy, especially at greater distances or with objects that are irregularly shaped. Integrating depth sensing with the current homography and size-based estimation approaches could reduce distance error margins and enhance the system's adaptability to larger and more complex environments.
- **Advanced Occlusion and Motion Blur Handling:** Machine learning techniques for marker detection could better handle occlusions, as these techniques can generalize and adapt to occluded shapes. Additionally, upgrading to a faster camera module or adding image stabilization algorithms may mitigate the impact of motion blur during quick rotations.
- **System Scalability and Modularity:** The current system design emphasizes modularity, making it relatively straightforward to integrate additional hardware or algorithms. Future iterations should continue to leverage this modular architecture, allowing for flexible upgrades and facilitating scalability to real warehouse environments. Further integration of real-time data analytics could also enhance task adaptability and system resilience in dynamic settings.

5.4 Concluding Remarks

The PicknPackRobotics TRL 3 prototype successfully demonstrates a foundational framework for autonomous warehouse robots, validating the feasibility of using a vision-centric approach for real-time

navigation, item detection, and retrieval. By combining computational efficiency, modularity, and scalable design, the system addresses core logistical challenges in warehouse automation, providing a glimpse into the future of robotics-driven supply chain operations. While improvements are necessary to ensure robustness across variable lighting, occlusions, and motion, the insights gained lay a strong groundwork for future research and development.

This project underscores the importance of balancing simplicity and performance in embedded systems, where hardware limitations can restrict processing power. Moving forward, the PicknPackRobotics team is well-positioned to refine this design for full-scale implementation, with the goal of revolutionizing warehouse automation through advanced computer vision and autonomous robotics.

References

- Agafonov, A. and Yumaganov, A. (2024). *IEEE Xplore Full-Text PDF:* [online] Ieee.org. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9253253 [Accessed 3 Nov. 2024].
- Australia Post. (2024). *Inside Australian eCommerce Report 2024 | AusPost*. [online] Available at: https://ecommerce-report.auspost.com.au/#authors-references [Accessed 3 Nov. 2024].
- Chen, D., Xu, K. and Ma, W. (2023). Binocular vision localization based on vision SLAM system with multi-sensor fusion. *2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, [online] pp.94–97. doi:https://doi.org/10.1109/cvidl58838.2023.10166820.
- Mamdouh, T. (2020). *Color spaces (RGB vs HSV) - Which one you should use?* [online] HubPages. Available at: https://discover.hubpages.com/technology/Color-spaces-RGB-vs-HSV-Which-one-to-use.
- Othman, N.A., Mehmet Umut Salur, Mehmet Karakose and Aydin, I. (2018). An Embedded Real-Time Object Detection and Measurement of its Size. [online] doi:https://doi.org/10.1109/idap.2018.8620812.
- Pankratova, A. (2024). *Computer Vision In Logistics And Warehousing*. [online] Opencv.ai. Available at: https://www.opencv.ai/blog/computer-vision-in-warehousing-and-logistics [Accessed 3 Nov. 2024].
- Pilarski, L., Luiz, L.E., Braun, J., Nakano, A.Y., Pinto, V., Costa, P. and Lima, J. (2023). An AI-based Object Detection Approach for Robotic Competitions. *Biblioteca Digital do IPB (Instituto Politecnico De Braganca)*, [online] 24, pp.1–6. doi:https://doi.org/10.1109/iceccme57830.2023.10252410.
- Sanjana Khan Shammi, Sultana, S., Islam, M.S. and Chakrabarty, A. (2018). Low Latency Image Processing of Transportation System Using Parallel Processing co-incident Multithreading (PPcM). *BRAC University Institutional Repository (BRAC University)*. [online] doi:https://doi.org/10.1109/iciev.2018.8640957.
- Sheryl (2023). *Mobile Picking vs Traditional Picking*. [online] Axacute. Available at: https://axacute.com/blog/mobile-picking-vs-traditional-picking/#:~:text=Traditional%20picking%20relies%20on%20paper-based%20systems%20or%20basic,through%20the%20warehouse%20to%20locate%20the%20required%20products. [Accessed 3 Nov. 2024].
- Taha, Rehman, Y., Rafiq, T., Nisar, M.Z., Ibrahim, M.S. and Usman, M. (2021). 3D Object Localization Using 2D Estimates for Computer Vision Applications. *arXiv (Cornell University)*. [online] doi:https://doi.org/10.1109/majicc53071.2021.9526270.