Monocular Human Detection using YOLO

Proposed By:
Navdeep Singh (120098024)
Sachin Jadhav (1194845224)

Product Overview

We propose integrating a human detection and tracking system into Acme Robotics' humanoid robot to enhance its perception capabilities in dynamic environments. This system will enable the robot to recognize and locate humans in various scenarios such as crowded indoor spaces or workplaces, facilitating tasks like guiding, assisting, and interacting with people. By utilizing detection and tracking data, the robot can adjust its actions, maintain safe distances, and prioritize human-focused tasks, improving its ability to operate autonomously and execute human-centered activities more effectively.

Humanoid robots increasingly perform human-centered tasks like customer service, healthcare, and collaborative work, which demand advanced perception capabilities for safe and effective interactions. Integrating a human detection and tracking system is critical for building a comprehensive perception stack, allowing the robot to identify and track humans, navigate complex environments safely, and respond to context. This specialized perception enables human-like spatial awareness, allowing the robot to avoid obstacles and perform socially aware tasks, enhancing user experience in real-world scenarios.

Definitions and Acronyms

YOLO (You Only Look Once): A state-of-the-art, real-time object detection system based on deep learning.

ONNX (Open Neural Network Exchange): An open format to represent machine learning models, facilitating the integration of pre-trained models across different frameworks.

NMS (Non-Maximum Suppression): A technique used to filter out redundant bounding boxes during the detection process, ensuring only the most confident predictions are kept.

CNN (Convolutional Neural Network): A class of deep neural networks widely used for analyzing visual data.

AIP (Agile Iterative Process): A development methodology characterized by iterative and incremental delivery of software features, enabling continuous improvement and adaptation.

Assumptions

Model Accuracy and Performance:

- It is assumed that the YOLO model, once converted to ONNX format, will retain a high level of accuracy in detecting humans and perform efficiently in real-time on the target hardware. The ONNX format is expected to provide optimal inference speed for the humanoid robot’s computational capabilities.

Environment Conditions:

- The system assumes reasonably controlled lighting conditions and minimal obstructions, as drastic changes in illumination or heavy occlusions may reduce detection accuracy. The humanoid robot is expected to operate primarily in environments where humans are clearly visible to its onboard camera(s).

Camera Quality and Configuration:

- The human detection system assumes that the robot's camera will provide high-quality images (sufficient resolution and frame rate) to enable reliable detection. It is also assumed that the camera will be positioned in a way that ensures a clear field of view for detecting humans at various distances.

NMS Threshold and Confidence Settings:

- Assumptions have been made regarding the appropriate Non-Maximum Suppression (NMS) threshold and confidence score settings that will yield a balance between detecting all humans in the scene and avoiding false positives. Fine-tuning may be required to adapt to specific application scenarios.

Hardware Compatibility:

- It is assumed that the humanoid robot's onboard processing unit is capable of running the YOLO-based detection algorithm without causing significant delays. The system will make use of available hardware acceleration (e.g., GPU or specialized inference hardware) if present.

Deliverables

The final deliverable will be a fully functional software application that uses real-time camera footage to detect, mark, and continuously update the position of humans relative to the camera, utilizing the user's webcam for demonstration. The software will operate efficiently in real-time with minimal latency. The source code will be rigorously tested using Google Test, achieving over 90% code coverage, and will adhere to the Google C++ style guide. Comprehensive documentation will be provided via Doxygen for all functions, classes, and modules to ensure clarity and ease of maintenance.

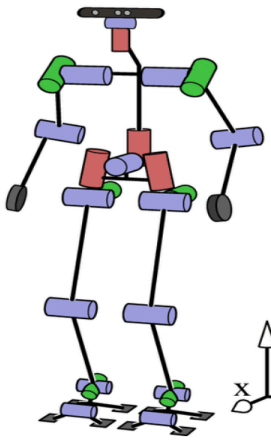
Cost

The primary costs associated with the project include computational resources for training and testing the detection and tracking models, as well as the labor associated with development, testing, and documentation. Since open-source libraries and pre-trained models are utilized, software costs will be minimized. However, additional expenses may arise if specialized hardware (e.g., GPUs) is required to optimize real-time performance on the robot.

Development Process and Organisation

The software development process will follow the Agile Iterative Process (AIP) with a pair programming approach. Sachin Jadhav will focus on developing the human detection component, while Navdeep Singh will concentrate on the human tracking system. To ensure quality, Sachin will be responsible for writing the unit tests, and Navdeep will execute these tests to validate the functionality. This sequence is chosen because the tracking system provides the final output of the perception stack. Both Sachin and Navdeep will contribute equally to the documentation for their respective components.

Diagrammatic Illustration



Joint	DoF
Head	2
Shoulder	2
Elbow	1
Pelvis	2
Hip	3
Knee	1
Ankle	2
Toe	1
Total	**24**

In the diagram, the degrees of freedom of the various joints of the robot are illustrated. The monocular camera will be attached to the head of the robot so as to increase its field of view enabling a better detection and subsequently a better functioning of the robot.