Danmarks Tekniske Universitet

Project Plan

DTU Bachelor Project

Nicklas Mundt
s224218

October 1, 2025

1 Introduction

Service robots are increasingly being deployed in human environments such as homes, hospitals, and public spaces, and their numbers continue to grow [1]. To operate effectively in human settings, robots must be able to perceive and understand their surroundings - a task that is natural for us humans but remains a central challenge for machines. In particular, the ability to recognize and localize objects is essential for meaningful interaction.

Traditionally, robots have relied on specialized sensors such as LiDAR or stereo cameras to achieve spatial understanding. While these approaches can deliver accurate depth information, they are often expensive or impractical for service robots. Vision-based methods provide an attractive alternative, as they infer spatial information directly from visual input in a way that resembles human perception. Among these, monocular depth estimation has gained attention for its ability to extract depth from a single image, thus also being a cheaper alternative. And with the introduction of Deep Neural networks, this field has really fostered over the last 10 years [2], with some of the first methods being explored in 2014 [3].

This bachelor project investigates how such vision-based methods can be applied to the humanoid Pepper robot. By combining object recognition with monocular depth estimation, the project aims to enable Pepper to localize and, if feasible, interact with physical objects in a controlled scenario.

2 Problem Formulation

Robots that interact with humans and physical environments require a robust understanding of their surroundings in order to perform tasks properly. A key element of this spatial understanding is the crucial ability to localize relevant objects within a given environment. Traditional approaches often rely on specialized sensors, such as LiDAR or stereo cameras, which may be expensive, limited in accessibility, or impractical in certain scenarios (for instance, autonomous cars usually uses LiDAR scanners for localization[4].

In contrast, vision-based approaches, particularly monocular depth estimation, offer a promising alternative by enabling depth perception from a single camera input[5]. Nonetheless, achieving accurate and reliable object localization through such methods remains a challenge, and ongoing research continues to propose new designs and techniques [2]. For service robots such as Pepper, which are intended to operate in dynamic human environments, selecting and implementing suitable methods for object recognition and spatial understanding is crucial to enabling meaningful interaction with their surroundings.

This project therefore addresses the following problem: How can vision-based methods, and specifically monocular depth estimation combined with object recognition, be applied to enable the Pepper robot to localize and interact with objects in a controlled scenario?

To address this, the project will be split up into three main parts:

1. *A survey* that will investigate different localization approaches, and the field of Depth Estimation, specifically within the monocular space, will be studied. It will explain the different methods used to achieve depth estimation, and will look at recent breakthroughs and methods in the field.
2. The implementation of an *object localization system.* This system will integrate Monocular Depth Estimation and Object Recognition to localize a specific object, and, if time allows it, let a Pepper robot move towards this object.

3. *An evaluation* of the system through practical experiments to assess its feasibility and performance in enabling spatial interaction.

In particular, the focus will be on enabling Pepper to recognize a specific object, estimate its spatial position through monocular depth estimation, and, if feasible, initiate simple interactions such as navigating toward the object or manipulating it.

To further structure the project, the following research questions are proposed for each part:

- **Survey**
  - What are the current state-of-the-art approaches for monocular depth estimation and vision-based object localization?
  - Which methods are most suitable for deployment on a service robot such as Pepper in a controlled environment?
- **Object localization system**
  - How can object recognition (e.g., using the powerful YOLO method[6],[7]) be effectively combined with monocular depth estimation to localize a target object?
  - Which technical challenges arise when integrating such a system on Pepper robots?
- **Evaluation**
  - How well does the system recognize the object?
  - To what extend can the Pepper interact with the environment?

These questions are subject to change as the project evolves.

3 Activity Plan

To keep a structured plan as well as a good overview of milestones, a Gannt chart was produced, as can be seen in figure 1. The chart illustrates the main phases of the project: The initial planning, the literature search for the survey, the implementation of the system and the report writing, which should include the evaluation of the system. As the process of structuring the project is inherently dynamic, the Gantt chart is expected to be adjusted as the work progresses. However, the overall structure will remain intact and guide the project timeline.

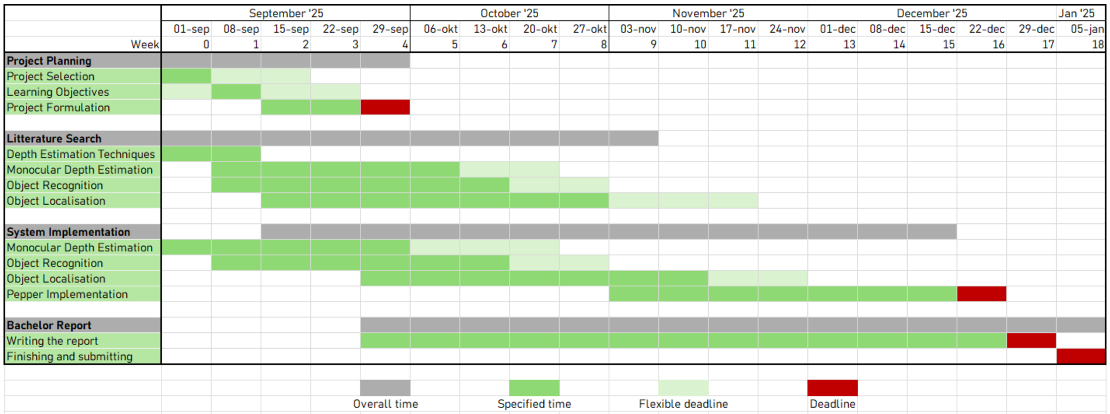

Figure 1: A rough Gantt chart visualizing the most important processes of the project, as well as some of the important deadlines for the project

Bibliography

[1] *Sales of Service Robots up 30 percent Worldwide*. 2024. URL: `https://ifr.org/ifr-press-releases/news/sales-of-service-robots-up-30-worldwide`.

[2] Zhen Xu et al. "Towards Depth Foundation Model: Recent Trends in Vision-Based Depth Estimation". In: *arXiv* (2025).

[3] David Eigen, Christian Puhrsch, and Rob Fergus. "Depth Map Prediction from a Single Image using a Multi-Scale Deep Network". In: *arXiv* (2014).

[4] Claudine Badue et al. "Self-Driving Cars: A Survey". In: *arXiv* (2019).

[5] *Monocular Depth Estimation*. 2023. URL: `https://huggingface.co/docs/transformers/main/tasks/monocular_depth_estimation`.

[6] Joseph Redmon et al. "You Only Look Once: Unified, Real-Time Object Detection". In: *The Computer Vision Foundation* (2016).

[7] Tianheng Cheng et al. "YOLO-World: Real-Time Open-Vocabulary Object Detection". In: *arXiv* (2024).