Imperial College
London

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Learning for SLAM in Dynamic Environments

Background Report

Author:
Ciro Cursio

Supervisor:
Stefan Leutenegger

Submitted in partial fulfillment of the requirements for the MSc degree in Computing (Machine Learning) of Imperial College London

June 2019

Contents

Chapter 1

Introduction

In recent times, computer vision applications have seen a dramatic rise in their popularity, from autonomous cars and robots to augmented reality systems. Their spatial awareness requirements are achieved by using Simultaneous Localisation and Mapping (SLAM), a strategy for jointly estimating a map of the environment and the position of the mapping agent in this environment: however, one of the biggest limitations of traditional SLAM algorithms is that the environment is assumed to be static, which limits the usefulness of this algorithm when the scene contains dynamic elements such as moving cars or people.

In the last years we have started to see SLAM systems that take into account such dynamics using strategies that can be roughly divided in three classes:

- **Rejection of Dynamic Elements:** this approach focuses on actively ignoring dynamic elements and only mapping the static parts of the scene.
- **Non-Rigid Object Reconstruction:** here the focus is on modelling only the dynamic parts of the scene, especially objects that deform in a non-rigid manner.
- **Multi-Object Tracking:** this approach combines both static and dynamic information, thus being able to create a map of the static parts of the environment, and simultaneously reconstruct and map moving objects that move rigidly.

The approaches that reject dynamic elements tend to have a high tracking accuracy of the camera pose, while approaches that reconstruct non-rigid objects can deal with extreme transformations. However, the strategy that has the largest potential and room for expansion is the one based on multi-object tracking: this approach tracks an arbitrary number of moving objects, thus allowing a greater degree of interaction between the agent and the environment (think of a robot manipulating objects in a room, or a virtual reality system that reacts when an object in the scene is moved).

At the core of such systems is a segmentation strategy that can distinguish between the camera pixels belonging to each moving object and the pixels belonging to the background. While the first dynamic SLAM systems focused on separating the static

parts and moving parts of the scene based on motion cues, in the last two years the focus shifted to semantic SLAM, where a deep learning-based segmentation system is used to identify each object and detect the region it occupies in the camera image. This presents a great advantage in the level of awareness that the agent gains (shifting from simply mapping the environment to understanding it), however it also introduces a significant speed limitation, since almost all segmentation networks are only optimised for accuracy and not speed: as a result, the state of the art systems for semantic instance segmentation are quite slow, operating at a maximum of 5 Hz on expensive high-end GPUs. This bottleneck is currently limiting dynamic SLAM systems in two aspects: first, it reduces their tracking capabilities in scenarios where the scene contain fast-moving elements, or when the agent itself performs rapid movements; second, it requires significant computational power, which in turn requires powerful hardware that is both expensive and cumbersome, especially on smaller robots and virtual reality headsets. This is the main problem that the present thesis aims to address: how can we improve the operating frequency of instance segmentation in the context of dynamic SLAM without a degradation in performance?

1.1 Aims and Objectives

The two main targets of this thesis are thus outlined below:

- **Development of a fast segmentation architecture:** following a literature review (included in this report) of the current methods for instance segmentation, the highest performing architectures in both accuracy and speed will be experimented with, and the insight gained will be used to build an architecture that reaches the target of an operating frequency of 30 Hz at the smallest possible degradation in segmentation accuracy compared to the state of the art.
- **Integration with an existing dynamic SLAM system:** in order to measure the real-world effectiveness of such a segmentation system, it will need to be integrated in a current dynamic SLAM system that employs moving object tracking, possibly one where the segmentation system is the computational bottleneck. After integration, the change in operating frequency, camera tracking and object reconstruction accuracy will be measured.

In addition, we define some additional targets that are not critical to the completion of the thesis, but present additional areas that could be interesting to explore:

- **Integration of depth information:** the most recent SLAM systems make use of RGBD cameras, which provide depth information together with RGB colour. However, the state of the art semantic segmentation methods only make use of RGB images, thus making it desirable for the SLAM system to further refine the segmentation masks using the scene depth in a subsequent and separate step.

Imperial College
London

Figure 1.1: Imperial College Logo. It's nice blue, and the font is quite stylish. But you can choose a different one if you don't like it.

An alternative could be to integrate the depth refinement in the segmentation network, which could result in a speed-up compared to having two separate steps.

- **Amodal instance semantic segmentation:** defined as the segmentation of objects *including the parts that are occluded by other objects*, amodal segmentation could improve object tracking in dynamic SLAM, as the object pose would be easier to estimate in difficult situations with many occlusions. However, amodal segmentation is an extremely recent research field, and thus the level of performance of the current methods is not as high as in the modal counterpart. This presents both a greater opportunity for original work, and a higher level of difficulty in creating a well-performing system.

Figure 1.1 is an example of a figure.

Chapter 2

Background

2.1 Semantic Segmentation

origins, encoder-decoder.

2.2 Instance Segmentation

Proposal-based methods (Faster RCNN -¿ MaskRCNN).
Oneshot methods. Semiconvolutional operators.

Chapter 3

Project Setup and Plan

Gantt chart here.
Also work packages.

Chapter 4

Conclusion