Training and Deploying Computer Vision Models for Indoor Localisation

Mika Senghaas
IT University of Copenhagen
jsen@itu.dk

A Thesis presented for the Degree of
Bachelor of Science in Data Science

IT UNIVERSITY OF CPH

IT University of Copenhagen
Computer Science Department

May, 15th 2023

Contents

1 Introduction **2**

2 Background **3**

3 Methodology **3**

3.1 Data Collection 4

3.2 Data Preprocessing 4

3.3 Models 5

3.4 Experiment Setup 5

4 Results **5**

5 Discussion **5**

6 Conclusion **5**

List of Tables

List of Figures

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

1 Introduction

Knowing where you are is crucial for human life. For as long as humans have lived, they have tried to determine their position on the earth. First, by observing the position of the sun, and later by using the stars. With advances in technology in the second half of the 20th century, specifically the invention and commercialisation of GPS (Global Positioning System), a satellite-based localisation system, localisation has become more efficient and accurate than ever before. Gradual commercialisation led to the technology rapidly transforming entire industries and personal navigation systems. Today, outdoor localisation is widely considered a *solved problem*.

The same cannot be said for indoor localisation. Because the transmitted radio signals sent out by the satellites in the GPS systems are not strong enough to penetrate through walls and struggle with reflections from large buildings, the technology yields inaccurate results at best, and often becomes dysfunctional in indoor spaces.

Finding alternative solutions to provide an accurate, cheap and robust indoor localisation systems has been a main focus of research in the past decades, and is becoming increasingly important in the light of the ongoing urbanisation of our living spaces and the emergence of autonomous robots and vehicles in our everyday life. Nonetheless, commercial applications are still rare and not unified in their approach.

Decades of research have led to the development of a variety of different indoor localisation technologies. Hardware-based systems use radio signals, transmitted by beacons, like Bluetooth, and Ultra-Wideband (UWB) or Wi-Fi, to localise an agent in a known environment. Software-based systems, like Simultaneous Localisation and Mapping (SLAM) algorithms, use sensors, like cameras or distance-measuring laser sensors, to localise an agent, while simultaneously creating a map of the environment.

However, hardware-based system require an expensive initial setup, continuous maintenance of the beacons, and are often not feasible in large environments, like shopping malls, or in environments that are frequently changing, like offices. SLAM algorithms, on the other hand, require a meticulously handcrafted pipeline of feature detection, feature matching, and pose estimation that has to be fine-tuned for each indoor space, to achieve outstanding results. Furthermore, some SLAM algorithms require specific types of sensors to be used, which are not available to use in all environments.

In an attempt to overcome the limitations of the aforementioned indoor localisation technologies, and to provide a simple, unified indoor localisation system this thesis proposes a novel approach to indoor localisation, by framing the problem of indoor localisation as a simple classification task. In our setup, location labels are continuously predicted from a stream of images, using different types of modern deep neural networks, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). With the advances in the computational power of modern mobile devices, the pipeline is proven to provide real-time estimates of the agent's position in the environment. Furthermore, the proposed method requires minimal initial setup in the environment and the devices, and is therefore suitable for commercial applications.

In this thesis we describe the data collection, model architecture, training procedure, and then rigorously evaluate the proposed method in three dimensions.

- **Accuracy:** Are the location estimates correct?
- **Robustness** Are the location estimates correct when encountering noise?
- **Efficiency:** How much data needs to be collected to train high-performance models? How quick is the inference times?

2 Background

Producing accurate, robust and cheap localisation systems is not a novel task, but has been a focus of research at the intersection of robotics, computer vision and machine learning for decades.

Amongst the most promising approaches are SLAM (Simultaneous Localisation and Mapping) algorithms. SLAM algorithms aim to localise an agent inside an unknown environment, while simultaneously building a consistent map of the environment. There exist a variety of different approaches to SLAM, depending on the type of sensors that are used for estimating position and mapping the environment. For example, Visual SLAM (V-SLAM) algorithms use camera input, and LidarSLAM algorithms use distance-measuring laser sensors. Initial proposal of such algorithms use a pipeline of feature detection, feature matching, and pose estimation to estimate the position of the agent and the environment.

The methodology most related to our approach are monocular V-SLAM algorithms, which use a single camera to estimate the position of the agent. The very first monocular feature-based V-SLAM algorithms is called MonoSLAM [1] and was proposed in 2007. The researchers proved that their approach is capable of simultaneously localising an agent and mapping an environment, using a single camera. Thus, overcoming the main challenge of inaccurate depth estimation using a single camera.

Since then, many adjustments and optimisation have been proposed to the algorithm to make it more robust and accurate. For example, the ORB-SLAM [3] algorithm uses a bag-of-words approach to feature matching, and the PTAM [2] algorithm uses a parallel tracking and mapping approach to improve the accuracy of the algorithm.

With machine and deep learning becoming more and more popular in the past decade, many researchers have started to apply these techniques to SLAM algorithms. For example, the DeepVO [4] algorithm uses a convolutional neural network to estimate the camera pose from a sequence of images. All the above mentioned approaches are similar to our approach in that they use deep learning at some part of the pipeline, but differ in that simply replace the original components in a SLAM pipeline, such as feature extraction, feature matching, or pose estimation, with a deep neural network, whereas we use deep learning to learn a mapping from images to location labels.

3 Methodology

The project is an end-to-end machine learning pipeline, meaning that it starts with the collection of data, and ends with the deployment of a production-ready model. The pipeline consists of the following steps, which will each be described in detail in the following sections.

1. Data Collection
2. Data Preprocessing
3. Different Model Architecture
4. Experiment Setup

3.1 Data Collection

By framing the problem of indoor localisation as a classification task, the project requires a labelled data set, which consists of pairs of images (frames in a video sequence) and location labels, (x_i, y_i), where x_i is a single frame of video footage, and y_i is the location label of the frame.

For the scope of this project, the data set was collected at the main building of the southern campus of the Copenhagen University (Danish: Københavns Universitet, KU) in Copenhagen, Denmark, and spans the two floors of the building. The location was chosen because it allowed for easy data collection and annotation, and because it is a relatively large building with similar indoor features, especially across floors, which was presumed to be a challenge for the model.

The data set was continuously collected from a single camera of a mobile device that was hand-held by a human agent while walking around the building. The camera was set to record video footage at the resolution of 2426x1125 pixels, and a frame rate of 30 frames per second (FPS). The camera was set to record video footage continuously, and the footage was stored on the device's internal storage, and later transferred to a computer for annotation and further processing.

The two floors in the building were separate into 21 different location labels following the building's floor plan. The location labels were denoted by descriptive identifiers. To match the location labels to the video footage, each video was manually annotated by denoting the starting and ending time stamps of a location label. The information was stored in a standardised format.

3.2 Data Preprocessing

The data set was pre-processed to make it suitable for fine-tuning several deep learning models. The pre-processing steps broadly involved resizing the video footage to a smaller resolution, down-sampling the frame rate, and splitting the data set into a training, validation and test set.

The video footage was resized to a resolution of 224x224 pixels, which is the input resolution of most modern foundation models for image and video classification, which were used in this project. Furthermore, it decreases the total data amount significantly, which allowed for less disk usage and faster loads into memory during training.

With a similar intention, the video was downsampled to a lower frame rate (between 1-5 FPS) to decrease the total data amount. It was hypothesised that because of the strong local dependency of consecutive frames, machine learning models would heavily overfit to the training footage. Empirical experiments confirmed that downsampling indeed reduces the over-fitting, and thus improve the robustness of the model against unseen data. The lower frame rate was therefore adopted for all further experiments.

Lastly, the video footage was split into a training, validation and test set, based on the date, on which the video footage was recorded. The training set only consists of data that was recorded on single days (February 22nd 2023 for the Ground Floor and March 3rd 2023 for the First Floor), while the validation and testing set consists of data that was recorded on multiple days (spanning from 9th-23rd of March 2023). This was a deliberate choice, in order to test whether an accurate model could be trained on only a single day of footage and still be robust against video clips with vastly different lighting conditions, temporal objects, and other factors that could affect the model's performance.

Some examples of frame-location label pairs after preprocessing are shown in Figure X.

3.3 Models

The project uses a fine-tuning approach to train various deep learning model for the task of indoor localisation. The fine-tuning approach is a common technique in machine learning, where a pre-trained model is used as a starting point for training a new model. The pre-trained model is usually trained on a large data set, such as ImageNet, and is therefore already capable of extracting useful features from images. The pre-trained model is then fine-tuned on a smaller data set, which is specific to the problem at hand. This approach is advantageous because it allows for the training of a model with a relatively small data set, while still achieving good performance.

3.4 Experiment Setup

Here I write something about the experiment setup, like the different research questions that I wish to answer and how I aim to test them. Each should list precisely which models are trained, under which hyper-parameters, and which metrics are used to evaluate the models.

1. Which model architecture performs best? (compare multiple different models and test with a set of metrics)
2. How does the model performance decrease when trained on smaller subsets of the data (This is important to assess how little the effort is to get a train an initial model that can be deployed)
3. How does the model performance decrease when trained on a single day of data?

4 Results

5 Discussion

6 Conclusion

References

[1] Andrew J. Davison, Ian D. Reid, Nicholas D. Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, 2007.

[2] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 225–234, 2007.

[3] Raul Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: a versatile and accurate monocular SLAM system. *CoRR*, abs/1502.00956, 2015.

[4] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. *CoRR*, abs/1709.08429, 2017.