ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Learned Single Shot Image-Based Camera Localization

Master Thesis

Moyuan Zhou

September 15, 2019

Advisors: Dr. Bugra Tekin, Dr. Johannes L. Schönberger

Supervisor: Prof. Marc Pollefeys

Department of Computer Science, ETH Zürich

Abstract

This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes.

Acknowledgements

Thanks Mum! This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes. This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes.

Contents

Chapter 1

Introduction

In recent years, image-based camera localization has been a key task in the areas of augmented reality, virtual reality, robotics, autonomous driving, etc. There are many methods relying on RGB-D cameras which are quite robust and accurate. However, the RGB-D cameras are power consuming, which makes approaches based on mobile and wearable cameras more attractive. In this thesis, we focus on camera localization from RGB images and aim for real-time, efficient, and robust camera localization.

Traditionally, the PnP [3] algorithm is used to solve the camera localization problem by computing the 6D camera pose given the 2D and 3D corresponding coordinates of multiple points. The fundamental problems of traditional approaches relying on local image features are textureless environments and robustness against strong changes in illumination/occlusion/viewpoint/etc. between the localized image and the given 3D model. To address these limitations, recently, deep learning has been applied to predict the 2D and 3D coordinates of these control points [1] and has achieved superior results. However, these approaches are typically very time-consuming to train and evaluate. In this thesis, we want to combine the benefits of both worlds and develop a learned approach that is efficient and robust.

The problem of 6D object pose estimation from RGB images is related to the problem of camera localization. Recently, an efficient, single shot approach [5] for simultaneously detecting an object in an RGB image and predicting its 6D pose without requiring multiple stages has been proposed. This approach resulted in an improvement over the state-of-the-art in terms of accuracy and efficiency, and addressed the challenges on keypoint occlusion and multiple object pose estimation. In this project, we aim to adapt this approach for efficient image-based camera localization. To this end, we will define keypoints on the 3D room layout for indoor environments and predict the projections of these 3D keypoints. The 6D pose will then be computed using PnP [3] based on the correspondences between 2D predictions and 3D

reference keypoints.

The major challenges of the project include limited data for training the localization task, occlusion, and motion. To address these issues, one initial idea is to increase the number of keypoints without slowing down the method, which is a direction to go for higher accuracy and robustness. [1] provided a trainable RANSAC approach for larger set of control points and can be integrated with the current model. Cutting edge methods could also be learned and utilized to achieve a higher performance.

1.1 Motivation

In recent years, image-based camera localization has great and wide applications in multiple areas. Autonomous localization and navigation is necessary for a moving robot. Augment reality on images requires camera pose or localization. To view virtual environment, corresponding viewing angle needs to be computed.

Furthermore, unlike some other technics that require special devices, e.g. Lidar sensors, RGB-D cameras, etc, cameras are ubiquitous nowadays and people carry with their mobile phones that have cameras every day. Thus, we want to utilize only RGB images from 2D cameras to realize image-based camera localization.

As the current approaches are time-consuming [1] or can not be generalized to new scenes [6], we aim to come up with an efficient, real-time, and robust camera localization approach.

1.2 Background

1.2.1 Overview on localization tools

There are many localization tools nowadays, among which GPS is very commonly used outdoors, but it cannot be used indoors. Lidar, UWB, WiFi AP et al are effective indoor localization tools. However, they require special devices or data collections in advance. Compared to these tools, camera photos can provide higher discriminated features and more information, but in the same time require higher computation ability.

There are also many effective and robust methods relying on depth information acquired by RGB-D cameras currently. However, the RGB-D cameras are power consuming and not ubiquitous as 2D cameras. Thus, passive RGB images that are more commonly used and easy to be acquired by mobile devices and wearable cameras become more attractive. In this work we focus on RGB images that could be captured by 2D cameras, and do not rely

on depth information. Compared to other localization tools, image-based camera localization is the most flexible and low cost one.

1.2.2 Image-based camera localization

Image-based camera localization [7] is to compute camera poses under some world coordinate system from images or video captured by the cameras. Image-based camera localization can be classified into two categories according to that environments are prior or not: the one with known environments and the other one with unknown environments. Then one with known environments are usually the PnP problem studies, and the one with unknown environments consists of the methods with online and real-time environment mapping and the methods without online and real-time environment mapping. The former is commonly known as Simultaneous Localization and Mapping (SLAM) and the latter is the middle procedure of the commonly known structure from motion (SFM). In this thesis, we are not doing any mapping or reconstruction for unknown environments since we are aiming for real-time localization given only single image as input.

There are also some approaches using convolutional neural network to predict camera pose directly from the 2D images or to compute 6D pose in some other way without using PnP algorithm. [6] predicts the orientation and translation of a camera given only a single picture. However, this approach is solving camera relocalization problem, and it can only predict in the same scene that the training period learnt. While this approach is useful in many robotic applications such as navigation and Simultaneous Localization and Mapping (SLAM), it cannot be generalised to a camera localization problem in a new/unseen scene. SSD-6D [2] relies on SSD architecture to predict 2D bounding boxes and a rough orientation estimate. Then based on the size of the 2D bounding box, it estimates the depth of the object and lift the 2D detection to 6D. However, the refinement step of this approach increases the running time a lot that cannot make a real-time prediction feasible.

Recently, a real-time single shot 6D object pose estimation approach [5] has been proposed. 6D object pose estimation is related with camera pose estimation which let us see a possibility of realizing real-time camera localization task in a general scene. [5] is using a single shot deep convolutional neural network to predict the 2D projections of the object's 3D bounding boxes, and then use a PnP algorithm to compute 6D object pose. We adapt this approach for our camera localization task. Namely, we use a PnP algorithm to calculate the camera pose from the 2D coordinates predicted by the network and the corresponding 3D keypoints of some known 3D models.

1.2.3 PnP algorithm

As we are applying PnP algorithm to calculate the camera pose according to some corresponding 2D and 3D points. Here is a introduction to PnP algorithm.

Camera pose determinations from known 3D space points are called perspective-n-point problem, namely PnP problem. Let n be the number of used points. When $n \geq 6$, the problem is linear. When $n = 3, 4, 5$, the problem is nonlinear. And when $n < 3$, there is no solution. Although the P3P problem has been well solved, but there may exist multiple solutions. In this work, we set the number of keypoints per object as 9. Although there may sometimes be outliers due to occlusion or other reasons, we can still guarantee it is a linear problem at most of the cases.

When the PnP problem is linear, there are also a lot of works studying on efficient optimizations for the camera poses from small number of points. [3] provide an accurate O(n) solution to the PnP problem, called EPnP which is widely used today. In our approach, we also utilize EPnP and compare it with other PnP solutions.

1.2.4 6D object pose estimation

There are also lots of current works on 6D object pose estimation. Traditional approaches are mainly local image features extraction and matching. There are many fast keypoint and edge-based methods for textured objects, but not effective to weakly textured or untextured objects, or low resolution video streams.

Recently, deep learning methods are utilized in solving 6D object pose estimation problem and have achieved outstanding performance. BB8 [4] is a 6D object detection pipeline made of one CNN to coarsely segment the object and another to predict the 2D locations of the projections of the object's 3D bounding box given the segmentation. It then use PnP algorithm to compute 6D pose using the corresponding points. The method is effective but slow due to its multi-stage nature. SSD-6D [2] is another approach which relies on SSD architecture to predict 2D bounding boxes and a very rough estimate of the object's orientation in a single step. Then the object's depth is approximated from the size of the 2D bounding box in the image and the 2D detection is lifted to 6D. However, both BB8 and SSD-6D methods require a post-processing step to refine the result, which increase the running time linearly and make the methods not feasible for real-time tasks.

1.2.5 Constraints on image-based camera localization problem

The common problem for camera localization or object pose detection is that the datasets available are rather limited, compared to datasets for other tasks like classification or tagging. Specifically, the most common datasets for detection contain thousands to hundreds of thousands of images with dozens to hundreds of tags, while classification datasets have millions of images with tens or hundreds of thousands of categories. Labelling images for detection is far more expensive than labelling for classification or tagging where tags are often user-supplied for free. These result in object detection problem lack of enough datasets and labels.

In this work, we generate the 2D projections of object's 3D bounding boxes for multiple objects as labels for indoor environments. We take Scannet dataset as the realistic dataset and SunCG together with six tools for synthetic dataset. The labels generated can also be used in future works on related topics.

1.3 Focus of this work

As clarified in the previous background section, we are aiming for a real-time and robust image-based camera localization approach that can predict the camera pose in a new scene. The new scene could be in similar environment as the training scenes, e.g. indoor environment with furnitures, but they should not be the same scenes. The approach should also be able to generalise to various object with no need of precise and detailed texture.

We adapt the recent real-time single shot 6D object pose prediction approach [5] to our camera localization approach. We also generate new labels for indoor environment dataset. Basically, the approach uses a single shot deep CNN architecture to find the corresponding 2D image coordinates of the 3D ground keypoints for some known 3D models in the environments and then applies a PnP algorithm to calculate the camera pose. We choose the 3D keypoints as the 8 corners of the object's 3D bounding box plus the center point of the object. The network is single shot and end-to-end trainable, and the model is also accurate without requiring any post-processing, so it is faster than the other current existing methods which are multi-stage or require post-processing for pose refinement.

Chapter 2

Related Work

In this chapter we review the existing works on camera localization from classical feature and template matching methods to newer end-to-end trainable CNN-based methods. Since we use 6D object pose estimation to realize our camera localization, object detection and 6D pose estimation related works are also reviewed.

2.1 Classical methods

Traditional RGB object instance recognition and pose estimation works are mainly local feature extraction and mapping.

2.2 RGB-D methods

2.3 CNN-based methods

2.3.1 Camera relocalization

2.3.2 Object detection

R-CNN and further methods propose various locations and regions

SSD

YOLO YOLO is a single shot object detection approach and is proposed in the same time as SSD. Most of the works for object detection before YOLO have a first step to propose possible regions of the object, and then do the classification and refinement. The multiple stages make the model hard to optimize since each component must be trained separately. The post-processing step also makes the training and prediction rather slow.

YOLO frame object detection as a regression problem instead of the previous classification problem. Instead of doing classification on a proposed region, YOLO can predict the location of the bounding boxes of the object directly. Given a full image, YOLO use a single network to predict all bounding boxes across all classes for the image simultaneously in one evaluation. Thus, YOLO realizes a single shot, end-to-end trainable network for object detection task.

Although YOLO cannot achieve the state-of-the-art accuracy but it is extremely fast and capable for real-time object detection. Compared with other real-time systems, it has more than twice of the mean precision. Although YOLO has lower recall compared to region proposal-based methods, it has less than half false positives on background errors compared to Fast R-CNN.

There are some constraints of YOLO as well. The first one is that there is a limit on the number of nearby objects that the model can predict. Since YOLO partition the image into multiple grid, and each grid cell can only predict one class, although it can predict B bounding boxes per grid, there is still only one object predicted as output. If two objects are very close and have their center in the same grid, only one of them can be predicted. The second one is that YOLO does not achieve the state-of-art accuracy. There is significant number of localization errors and relatively low recall of YOLO compared to region proposal-based methods.

YOLO v2 YOLO v2 has done some improvements to the YOLO detection method. Inspired by Faster R-CNN, which use the region proposal network(RPN) to predict offsets and confidences for anchor boxes, YOLO v2 removed the fully connected layer at the end of the network, and instead use anchor boxes to predict bounding boxes. YOLO v2 decouples the class prediction mechanism from the spatial location and instead predict class for every anchor box. In this way, the limitation of only one class is predicted per grid cell is released, and now we can better predict nearby objects.

YOLO v2 achieves a good tradeoff between speed and accuracy. It outperforms Faster R-CNN with ResNet and SSD while still running significantly faster. In this work, our network architecture is also based on YOLO v2.

2.3.3 6D object pose detection

BB8 BB8 [4] is a 6D object detection approach which consists of one CNN to realize a coarse segmentation and another to predict the 2D locations of the projections of the object's 3D bounding box, which are then used by a PnP algorithm to compute the 6D object pose. The segmentation stage generates a 2D segmentation mask for presenting a cropped image to the second network. There is also an optional additional step that refines the predicted poses. The method is slow due to its multi-stage nature.

SSD-6D The SSD-6D [2] approach relies on the SSD architecture to predict object's 2D bounding boxes and a pool of the most likely 6D poses for that instance. It then predicts the approximated depth of the object from the size of the 2D bounding box in the image and lift the output from 2D to 6D object pose. In the final step, the approach refine each pose in every pool and select the best after verification. As the method require a refinement step to get a good accuracy, the running time is increased linearly with the number of objects being detected.

Real-time single shot approach The recently proposed real-time seamless single shot approach [5] performs 6D object pose prediction in an RGB image without multiple stages or hypotheses. It proposed a new CNN architecture which is inspired by YOLO and YOLO v2. The end-to-end trainable architecture makes it easy and fast to optimize. The approach turns out to be fast enough for real-time 6D object pose detection and accurate without requiring any additional post-processing.

The network is based on the YOLO v2 network but extends 2D detection to 6D detection task. It directly predicts the 2D image locations of the projected vertices of the object's 3D bounding box and then use a PnP algorithm to predict the object's 6D pose.

Compared with other approaches, the single shot approach outperforms BB8 [4] and SSD-6D [2] when they are tested without post-processing - as we introduced above, BB8 and SSD-6D have pose refinement step as post-processing to boost the accuracy but with a cost of much slower performance. When handling multiple objects, the single shot approach virtually has no time-penalty, that is the running time remains constant, whereas other methods grow proportional to the number of objects.

Chapter 3

Methodology

This work is following B. Tekin's real-time single shot 6D object pose detection method [5] which is inspired by the performance of YOLO and YOLO v2 on single shot 2D object detection. The network architecture is based on YOLO v2 but extends 2D detection to 6D detection task. The output of the network is the 2D coordinates of the projections of the 3D keypoints of the object. The approach then applies a PnP algorithm to compute the camera pose according to the corresponding 2D and 3D points.

In this section, we first review the network architecture of YOLO and YOLO v2 in details, and then elaborate our improvements on it for a camera localization problem.

3.1   YOLO and YOLO v2

YOLO [?] is a single shot object detection approach which first frames object detection as a regression problem to spatially separated bounding boxes and associated class probabilities, instead of the prior repurposing to classification works.

YOLO uses a single neural network to predict bounding boxes and class probabilities directly from full images in one evaluation. Given an input image, the system divides the image into an $S \times S$ grid. The grid that the center of an object falls into is responsible for detecting that object. Each grid cell predicts B bounding boxes and confidence scores for those boxes. The confidence score reflects how confident the model is that the box contains an object and also how accurate it thinks the box that it predicts is. YOLO takes the intersection over union (IOU) as the measurements for the confidence score. Formally, the confidence is defined as $Pr(Object) * IOU_{pred}^{truth}$. That is, if no object exists in the cell, the confidence is 0. Otherwise the confidence score equals the intersection over union (IOU) between the predicted

box and the ground truth. Each grid cell also predicts C conditional class probabilities, $Pr(Class_i|Object)$. These probabilities are conditioned on the grid cell containing an object. There is only one set of class probabilities predicted per grid cell, regardless the number of bounding boxes B is predicted per cell.

3.2 Model

3.3 Training

3.4 Prediction

3.4.1 Example Subsection

Dummy text.

Example Subsubsection

Dummy text.

Example Paragraph Dummy text.

Example Subparagraph Dummy text.

Chapter 4

Datasets

Dummy text.

4.1 Dataset types

Dummy text.

4.1.1 Synthetic data

Dummy text.

4.1.2 Realistic data

Dummy text.

Scannet

Example Paragraph Dummy text.

Example Subparagraph Dummy text.

Chapter 5

Experiments and Results

Dummy text.

5.1 Example Section

Dummy text.

5.1.1 Example Subsection

Dummy text.

Example Subsubsection

Dummy text.

Example Paragraph   Dummy text.

Example Subparagraph   Dummy text.

Chapter 6

Constraints and Future Work

Dummy text.

6.1 Example Section

Dummy text.

6.1.1 Example Subsection

Dummy text.

Example Subsubsection

Dummy text.

Example Paragraph   Dummy text.

Example Subparagraph   Dummy text.

Appendix A

Dummy Appendix

You can defer lengthy calculations that would otherwise only interrupt the flow of your thesis to an appendix.

Bibliography

[1] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6684–6692, 2017.

[2] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1521–1529, 2017.

[3] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, 81(2):155, 2009.

[4] Mahdi Rad and Vincent Lepetit. Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3828–3836, 2017.

[5] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 292–301, 2018.

[6] Jian Wu, Liwei Ma, and Xiaolin Hu. Delving deeper into convolutional neural networks for camera relocalization. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5644–5651. IEEE, 2017.

[7] Yihong Wu, Fulin Tang, and Heping Li. Image-based camera localization: an overview. *Visual Computing for Industry, Biomedicine, and Art*, 1(1):1–13, 2018.

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

_____

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):   **First name(s):**

With my signature I confirm that
   − I have committed none of the forms of plagiarism described in the '<u>Citation etiquette</u>' information sheet.
   − I have documented all methods, data and processes truthfully.
   − I have not manipulated any data.
   − I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date   **Signature(s)**

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.