AIVERO AS Internship

Group 18gr942

Project Report

Spring Semester 2018

Aalborg University
Vision, Graphics and Interactive Systems

This report is compiled in LaTeX. Additionally is Mathworks MATLAB, Python, Adobe Illustrator, and Inkscape used to code, draw figures, and charts.

Vision, Graphics and Interactive Systems
Aalborg University
http://www.aau.dk

AALBORG UNIVERSITY
STUDENT REPORT

Title:
AIVERO AS Internship

Theme:
Computer Vision

Project Period:
Spring Semester 2018

Project Group:
Group 18gr942

Participants:
Niclas Hjorth Stjernholm

Supervisor:
Thomas B. Moeslund
Raphael Düershceid

Number of Pages: 22

Date of Completion:
January 9, 2018

Abstract:

Preface

This report is composed by Niclas Stjernholm during the third semester of the master's programme in Vision, Graphics and Interactive Systems at Aalborg University. The third semester is used as company internship.

For citation the report employs the Harvard method. If citation is not present in tables or figures, they are produced by the author.

This project is implemented in MATLAB 2017a and Python 3.6.

Aalborg University, January 4, 2019

Niclas Hjorth Stjernholm
<nstjer14@student.aau.dk>

Contents

Glossary

CNN	Convolutional Neural Network.
FC	Fully Connected.
GQ-CNN	Grasp Quality Convolutional Neural Network.
MOT	Multiple Object Tracking benchmark.
SVM	support vector machines.
YOLO	You Only Look Once.

Chapter 1

Introduction

Several warehouses around the world are starting to use robots for picking groceries of different sorts [Olsen, 2018; Perez, 2018; Vincent, 2018]. This is done to automate the process of grocery picking and make the entire process faster, to be able to deliver packages faster. Olsen [2018]; Perez [2018]; Vincent [2018] present robots picking from boxes and not free standing products like in a general supermarket.

The goal of Aivero is to enable robots in a general supermarket set-up to pick groceries straight from the shelves with only the information of which section the desired grocery is. The aim is do this using a video stream directly from the robot and process this data in real time. Using the video stream a processing unit must identify the desired grocery and find the optimal picking point on the grocery. For a potential higher accuracy the goal is to use both a regular colour video stream, RGB, and depth video. This will enable the robot to see the groceries in physical shape better from the depth video, but also use th RGB video stream to recognise e.g. labels on a product.

Chapter 2

Background Research

As a part of the implementation of a viable grasping point of objects in an environmental setting a research of state of the solutions of both object tracking and detection is necessary as well as researching full existing solutions.

2.1 Object Detection and Tracking

In order to get some understanding of the state of the art technologies used the research in split into multiple sections for different areas in the process of teaching a robot to pick groceries from a shelf.

2.1.1 Object Detection

The aim of generic object detection is to locate and classify an object in an image, sometimes including a bounding box on objects with a confidence of existence in the image. In general there are two approaches to object detection namely region proposal, which follows the traditional object detection pipeline and then classifying the each proposal into different object categories. The other approach is the regression or classification based approach by adopting a unified framework to achieve final results [Zhao et al., 2018]. Zhao et al. [2018] presents several solutions of both approaches and have also made a small roadmap of different popular solutions up until 2017. This is shown in Figure 2.1.

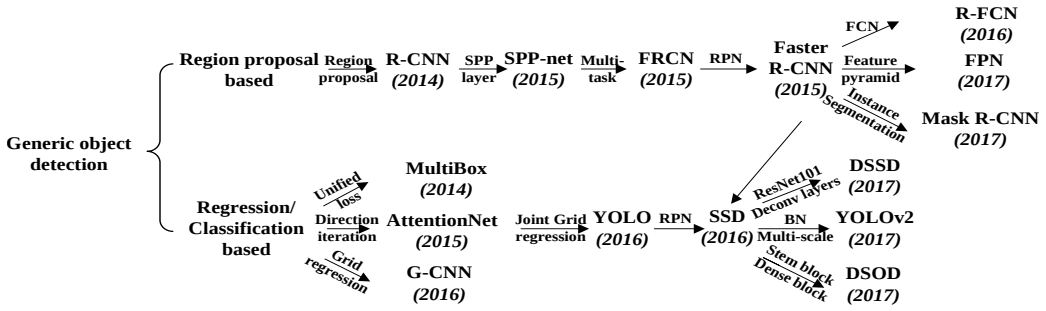

Figure 2.1: Small roadmap of different popular solutions up until 2017 [Zhao et al., 2018]

Another way of doing object detection is using decision trees instead of deep learning as proposed by Gall et al. [2012].

Region Proposal

The region proposal framework is a two step process. The framework firstly scans an image and then focuses in any regions in the image of interest. As shown in Figure 2.1 the R-CNN solution is one of the first in region proposal approach. The design has three stages in the process of object detection. Firstly a region proposal generation, generating 2000 region proposals for each image. Afterwards a Convolutional Neural Network (CNN) is applied to extract features from warped or cropped regions, extracting a 4096-dimensional feature vector. From there classification and categorisation is done with pre-trained support vector machines (SVM)s from multiple classes. The final bounding boxes are produced from adjusted scored regions using bounding box regression.

Regression or Classification framework

To compete with the time consumption of regional proposal the one-step framework based on global regression/classification is utilised by mapping straight from image pixels to bounding box coordinates and class probabilities. This is the technique the framework You Only Look Once (YOLO) applies. YOLO divides an input image into $S \times S$ grid, each cell is responsible for predicting the object centred in the cell. The first YOLO framework has 24 convolutional layers and 2 Fully Connected (FC) layers [Zhao et al., 2018]. The second version, referred to as both YOLOv2 and YOLO9000, uses the Darknet-19 model, which has 19 convolutional layers and 5 maxpooling layers. It is based on the Googlenet architecture [Redmon and Farhadi, 2016]. At the point of deployment this framework has state of the art performance. The third iteration of the framework YOLOv3 with an update of the Darknet architecture increasing the size from 19 convolutional layers to 53. The newest YOLO framework is from early 2018 and is still one of the best in its category. In precision Redmon and Farhadi [2018] states the framework does not perform as well as RetinaNet, but the speed of the framework is faster. YOLO is fast and lightweight enough to run

in real time. It is mostly trained on the COCO and VOC datasets [Redmon and Farhadi, 2018].

Decision Trees

A decision tree consists of split nodes and leaves. The split nodes evaluate each image patch based on a set decision metric and pass the patch to either left or right in the tree. The leaf is the end of a tree and stores the statistics of the image patches which arrived during training of the tree. This is illustrated in Figure 2.2.

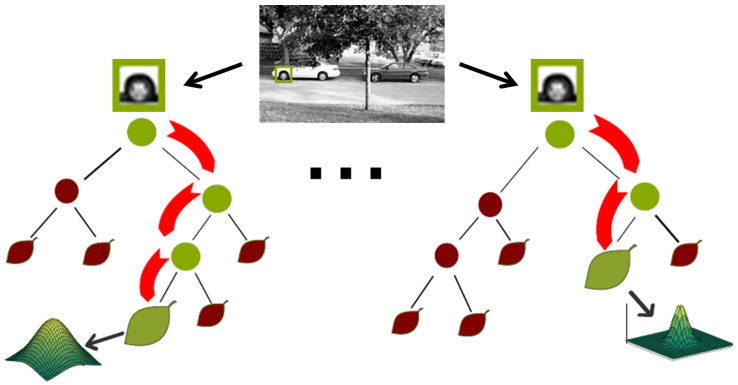

Figure 2.2: Example of decision tree evaluation on image patch from bounding box [Gall et al., 2012]

Gall et al. [2012] uses a random forest for multi class object detection. A random forest consist of a set amount of decision trees. The detection problem becomes a distribution estimation problem, where the random forests allow to learn features and descriptors that are optimal for estimating the distributions with low uncertainty. These distributions may be desirable to restrict to Gaussians or Gaussian Mixture Models [Gall et al., 2012].

The amount of training data is the most crucial parameter for detection accuracy, as random forests are designed to handle large amounts of data. If it is small training sets extra verification steps may be needed [Gall et al., 2012].

2.1.2 Tracking

General object tracking is tracking an object in a video or a sequence of images, in form of e.g. the problem of estimating the trajectory of an object in the image plane as the object moves in the image sequence. This means, that the tracker consistently labels the tracked objects in the different frames of the video [Acm Reference Format: Yilmaz et al., 2006]. Situations often present tracking of more than one object at the time where a multi object tracking framework is needed. Leal-Taixé et al. [2017] has evaluated on 32 different state of the art trackers from no sooner than March 2017 on multiple objects. The evaluation is performed the Multiple Object Tracking

benchmark (MOT)15 and MOT16 datasets. This is done to introduce a standardised benchmark and analyse the performance of the trackers [Leal-Taixé et al., 2017]. Leal-Taixé et al. [2017] states there are six top-performing trackers with a multiple object tracking accuracy higher than 40%. Of the top six trackers, two of them are using deep learning, one using Recurrent neural networks to encode appearance of pedestrians and the other is using deep matching. The common component of the top performers are strong affinity models [Leal-Taixé et al., 2017].

Three Dimensional Object Tracking

In order to utilise three dimensional video, another dimension needs to be added by introducing depth in the data. This is done by using a depth camera to record and colouring the video in accordance to depth.

Tan and Ilic [2014] propose a solution for both two dimensional template an three dimensional object tracking using multi forest decision trees. This is based on another objective function which relates the image intensities of a template and transformation parameters by the pseudo-inverse of a Jacobian matrix, but with the use of random forests instead of the Jacobian, making the method generalised to any input function and not constrained to two dimensional intensity images [Tan and Ilic, 2014]. The random forests are regression forests to learn how different values of the transformation function affect the input function. The training of forests begins by creating a training dataset with n_ω random values to transform the input to affect the computation of the sample points. They randomly select n_r points from n_s sample points of the template for constructing a tree to impose randomness. Tan and Ilic [2014] uses 100 trees for each 6 parameters with a grid enclosed on the model seen from the depth image, only the points which lie on the model are used as sample points. In training 50 000 depth images are rendered with different kinds of distortions, such as rotation and translation. This is done for each camera view, of which there are 42.

Object Pose

For a robot to be able to interact with an object, more information than the position in an image is needed. The orientation of the object is also important. Xiang et al. [2017] seeks to estimate the six degrees object pose with the network, PoseCNN. PoseCNN, is trained using RGB-D scans of 21 objects and evaluated on both the YCB dataset and the OccludedLINEMOD dataset, to evaluate the robustness towards occlusions. The network has three different tasks done in order to estimate a six dimensional pose. Firstly a semantic labelling, secondly a 3D translation estimation, and thirdly a 3D rotation regression [Xiang et al., 2017]. This is illustrated in Figure 2.3.

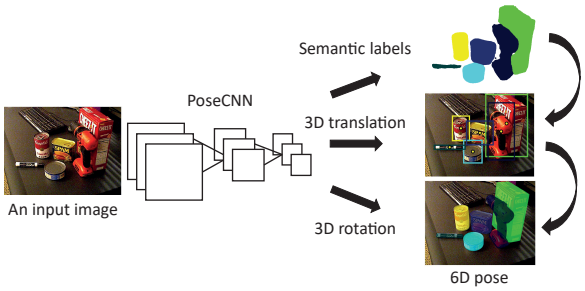

Figure 2.3: Overview of the three steps performed in by the PoseCNN [Xiang et al., 2017]

2.1.3 Grasping Planning

For a robot to interact with an object a grasping point on the tracked object is necessary. Mahler et al. [2017a] proposes a solution of grasping planning in two different solutions based on the same deep learning model called Grasp Quality Convolutional Neural Network (GQ-CNN). They have created their own dataset with 1 500 3D object models fro training the network. Before training of the network for the second iteration Dex-Net 2.0, Mahler et al. [2017a] analyse a dataset of 6.7 million point clouds with grasping options which has been generated from 3D models in randomised poses on a table. For the third iteration, Dex-Net 3.0 is generated. 2.8 million point clouds with suction grasps, and grasps robustness labels on 1 500 3D objects models. The database is used to train the GQ-CNN to classify suction grasp robustness. This is done with a model of the pneumatic suction gripper used with the robot when testing [Mahler et al., 2017b].

Dex-Net 2.0 is tested 3D printed objects designed to challenge the parallel gripper used in this iteration and 10 household objects [Mahler et al., 2017a]. Dex-Net 3.0 is tested on objects divided into three different categories of basic, typical, and adversarial. Both the object categories and the performance in the different groups are shown in Figure 2.4.

	Basic		Typical		Adversarial		
	AP (%)	Success Rate (%)	AP (%)	Success Rate (%)	AP (%)	Success Rate (	
Planarity	81	74	69	67	48	47	
Centroid	89	92	80	78	47	38	%)
Planarity-Centroid	98	94	94	**86**	64	62	
GQ-CNN (ADV)	83	77	75	67	**86**	**81**	
GQ-CNN (DN3)	**99**	**98**	**97**	82	61	58	

Figure 2.4: Dex-Net 3.0 object categories and performance. Basic (e.g. prismatic objects), Typical, and Adversarial [Mahler et al., 2017b]

Chapter 3

Project Specification

This chapter specifies the scope of the project. It outlines and delimits the goals for the work conducted, as well as setting the requirements for the solutions implemented during the project work.

The two objectives, first one set as an open objective the second closed, given by the company was originally:

- Improve depth quality of commodity RGB-D cameras
- 6 DOF object pose tracking of known objects using 3D cameras (for robotic picking of groceries)

The primary objective was the second and closed objective, to implement a 6 DOF object pose tracking. As a part of smaller company other tasks also needed work, which took time from the amount of time for the primary objective.
Other tasks were:

- Train a YOLOv3 network for object detection in a Python implementation
- Gstreamer command testing and configuration
- Setting up a docker for testing
- Create small test setup for matrix storage in C++

Working on these other tasks meant the main task got scaled down due to time constraints. Instead, the first steps of the object pose tracking became essential for showing potential in the idea. This mainly focused on finding an already working solution of object detection and grasping point and showing a real time object detection solution for groceries.

Bibliography

Acm Reference Format: Yilmaz, A., Javed, O., and Shah, M., 2006. Object tracking: A survey. *ACM Comput. Surv*, 38, p. 45. doi: 10.1145/1177352.1177355. Available at: <http://doi.acm.org/10.1145/1177352.1177355>.

Gall, J., Razavi, N., and Van Gool, L. An introduction to random forests for multi-class object detection. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7474 LNCS, pp. 243–263, 2012. ISBN 9783642340901. doi: 10.1007/978-3-642-34091-8_11. Available at: <http://link.springer.com/10.1007/978-3-642-34091-8{_}11>.

Leal-Taixé, L., Milan, A., Schindler, K., Cremers, D., Reid, I., and Roth, S., 2017. Tracking the Trackers: An Analysis of the State of the Art in Multiple Object Tracking. apr. Available at: <https://arxiv.org/abs/1704.02781>.

Mahler, J., Liang, J., Niyaz, S., Laskey, M., Doan, R., Liu, X., Ojea, J. A., and Goldberg, K., 2017. Dex-Net 2.0: Deep Learning to Plan Robust Grasps with Synthetic Point Clouds and Analytic Grasp Metrics. mar. ISSN 0929-5593. doi: 10.15607/RSS.2017.XIII.058. Available at: <http://arxiv.org/abs/1703.09312>.

Mahler, J., Matl, M., Liu, X., Li, A., Gealy, D., and Goldberg, K., 2017. Dex-Net 3.0: Computing Robust Robot Vacuum Suction Grasp Targets in Point Clouds using a New Analytic Model and Deep Learning. sep. doi: 10.1109/ICRA.2018.8460887. Available at: <http://arxiv.org/abs/1709.06670>.

Olsen, L., nov 2018. *Fuldautomatisk robotlager: Nu tager det syv minutter at pakke en pakke*, Horsens. Available at: <https://hsfo.dk/erhverv/Fuldautomatisk-robotlager-Nu-tager-det-syv-minutter-at-pakke-en-pakke/artikel/199254>.

Perez, S., 2018. *Walmart pilots a grocery-picking robot to fulfill customers' online orders*. Available at: <https://techcrunch.com/2018/08/03/walmart-pilots-a-grocery-picking-robot-to-fulfill-customers-online-orders/?guccounter=1>.

Redmon, J. and Farhadi, A., 2016. YOLO9000: Better, Faster, Stronger. dec. Available at: <http://arxiv.org/abs/1612.08242>.

Redmon, J. and Farhadi, A., 2018. YOLOv3: An Incremental Improvement. apr. Available at: <http://arxiv.org/abs/1804.02767>.

Tan, D. J. and Ilic, S. Multi-forest tracker: A Chameleon in tracking. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1202–1209. IEEE, jun 2014. ISBN 9781479951178. doi: 10.1109/CVPR.2014.157. Available at: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6909553>.

Vincent, J., 2018. *WELCOME TO THE AUTOMATED WAREHOUSE OF THE FUTURE*. Available at: <https://www.theverge.com/2018/5/8/17331250/automated-warehouses-jobs-ocado-andover-amazon>.

Xiang, Y., Schmidt, T., Narayanan, V., and Fox, D., 2017. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. nov. Available at: <https://arxiv.org/abs/1711.00199>.

Zhao, Z.-Q., Zheng, P., Xu, S.-T., and Wu, X., 2018. Object Detection with Deep Learning: A Review. Available at: <https://arxiv.org/pdf/1807.05511.pdf>.