

Imperial College London

Department of Electrical and Electronic Engineering

Final Year Project Report 2019

Project Title: **Augmented Reality-assisted Human Robotic Interaction**

Student: **Aufar P. Laksana**

CID: **01093575**

Course: **EIE4**

Project Supervisor: **Dr Yiannis Demiris**

Second Marker: **Dr Tae-Kyun Kim**

Powered wheelchairs are becoming increasingly commonplace in the modern world. However, a major issue faced by powered wheelchair users (PWUs) is navigating the device in crowded areas. Controlling the powered wheelchair in crowded areas requires increased concentration from the PWU, as people in crowds often move unpredictably, or are hidden from view due to standing behind another person or object.

This project utilizes computer vision techniques to predict the direction of travel of individuals in crowds, and implements an augmented reality system using the Microsoft HoloLens that aids the PWU by displaying visual aids that indicate the motion of people. The system further aids the user by warning the user of potential collisions, allowing the PWU to make better navigation decisions. The project also explores the use of the system as a method of assistive control of the wheelchair, preventing collisions by stopping the powered wheelchair should the PWU not notice an individual crossing their path.

Contents

1	Introduction and Requirements	3
1.1	Introduction	3
1.2	Motivation	3
2	Background	4
2.1	Human Detection	4
2.1.1	Definition of Requirements	4
2.1.2	Review of Existing Methodologies	4
2.1.3	Comments	8
2.2	Object Tracking	8
2.3	Head Pose Estimation	8
2.4	Body Pose Estimation	8
2.5	SLAM	8
2.6	Augmented Reality Headsets	8
3	Requirements Capture	9

Chapter 1

Introduction and Requirements

1.1 Introduction

This report was written as part of the Final Year Project for the MEng Electronic & Information Engineering course. The project was supervised by Dr. Yiannis Demiris at the Imperial College London.

1.2 Motivation

Chapter 2

Background

This project is focused on computer vision for detecting and tracking humans in the surroundings, estimating their trajectories and distance from the PWU, the reactive control systems that prevent collisions with the detected objects as well as the augmented reality display to provide visual cues to the PWU.

2.1 Human Detection

Human detection is a subset of the classic computer vision problem of object detection. In order to develop an augmented reality system that will help PWUs to navigate in public spaces, it is essential for the system to be able to discern humans from the surroundings.

2.1.1 Definition of Requirements

The problem arises in crowded areas, whereby individuals are occluded by other people or objects in front of them, leaving only certain body parts visible. As such, we began our research with the problem of being able to detect people in images where identifying parts of the body are not always visible.

2.1.2 Review of Existing Methodologies

A related field of research is that of people counting and human detection in visual surveillance in public areas. Where the problem differs is that surveillance benefits from being able to rely on cameras with a good view of the crowd from above, whereas for a PWU, the camera will not have as high of a vantage point, making detecting every single individual in a crowd impossible.

Despite the disadvantage, similar techniques can be used to detect humans in video. Most methods can be classified into two categories [1]. The first technique, foreground detection, attempts to model the background of an image and then detect the changes that occur between frames. The second category involves exhaustively searching the image with a scanning window, and deciding if each window can be classified into a human shape.

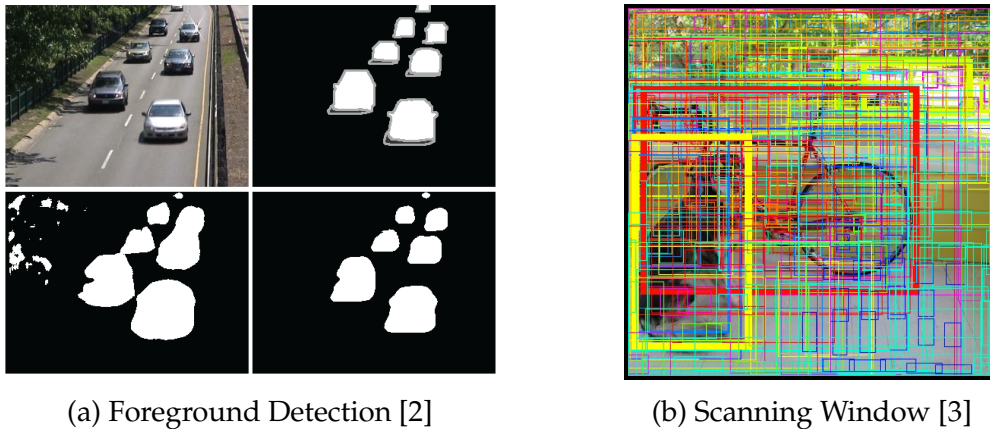


Figure 2.1: Comparison of Foreground Detection and Scanning Windows

Foreground Detection

Background subtraction is a widely used approach for detecting moving objects [4]. A temporal average filter can be used to find the median of all the pixels in an image to form a reference image. Frames with moving objects can then be compared pixelwise to the reference, and a threshold set to determine if the pixel is part of the background or foreground. People counting and human detection can then be achieved by segmenting the foreground image into individuals.

However, this technique relies on a static camera in a well placed location. This brings up several reasons as to why this method would not be suitable for this project. Firstly, the camera available is part of a head-mounted augmented reality device. The wearer has the ability to move the camera in 6 degrees of freedom. Secondly, the wearer will also be navigating a powered wheelchair. As a result, the background is constantly changing, and the reference image would require constant recomputation before human detection can even begin.

Scanning Windows

Due to the ever-changing surroundings of a mobile robot, a better approach for object detection is to exhaustively search an image using scanning windows and determining if an object was detected in each window. However, it must be noted that this method is computationally expensive. In order to achieve real-time detection on a mobile robot, the use of a graphics processing unit (GPU) should be considered [5].

Classical Object Detection

Haar Cascades Haar cascades classifies images based on the value of simple features [6], which are variants of the difference between the sum of pixel values in rectangular regions. An intermediate representation of the original image is used to rapidly compute a small set of representative rectangular features.

A cascade of classifiers is then used to determine if the region is detected as a human. The detection process is that of a degenerate decision tree, where a positive result in the first cascade will trigger an evaluation in the second, more successful classifier. As such, the initial classifier can eliminate a large number of negative examples with very little processing. After several stages, the number of sub-windows has been reduced radically

Histograms of Oriented Gradients The method proposed is implemented by dividing the image window into small spatial regions and calculating a local 1-D histogram of gradient directions for all the pixels in the region. The combined local histograms form the overall feature representation of the image.

The detection window is tiled with the Histogram of Oriented Gradient (HOG) descriptors. In the original paper [7], these feature vectors were then used in a conventional SVM based window classifier to give human detections.

Deep Learning Object Detection

Modern approaches for human detection largely depend on Deep Convolutional Neural Networks (CNN). The approach provides the best in class performance, as well as scaling effectively with more data. An added advantage of using CNN based object detection systems for this project is that they are also capable of detecting multiple classes of objects.

An issue with CNN approaches is that the methods are trying to draw bounding boxes around objects of interest in images. However, we do not know the number of objects in the image beforehand. As such, to be completely sure every object has been detected, a naive solution is to take a huge number of regions and attempt to classify all the objects in the region, a computationally expensive process.

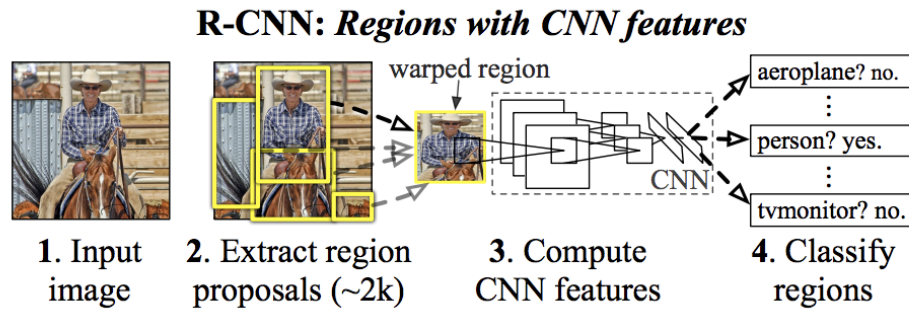


Figure 2.2: R-CNN Approach

R-CNN The R-CNN method uses a selective search to extract 2000 regions from an image [8]. The regions are selected by generating a large number of candidate regions and using a greedy algorithm to recursively combine similar regions into larger ones. The regions are then fed into a CNN that acts as a feature extractor and the output dense layer consists of the features extracted from the image, which are then fed into an SVM to classify the presence of objects in the region.

The major disadvantage to this approach is the amount of time required to train the network. Each training image has to be classified once for each of the 2000 region proposals. Furthermore, the selective search algorithm is a fixed algorithm (no learning is done), and as such, could lead to generation of bad candidate region proposals.

YOLO Whereas R-CNN uses regions to localize the object within an image, You Only Look Once (YOLO) looks at the image as a whole and uses a single CNN to predict the bounding box and the class probabilities [3]. By looking at the image as a whole, the network can use features from the entire image to predict each bounding box.

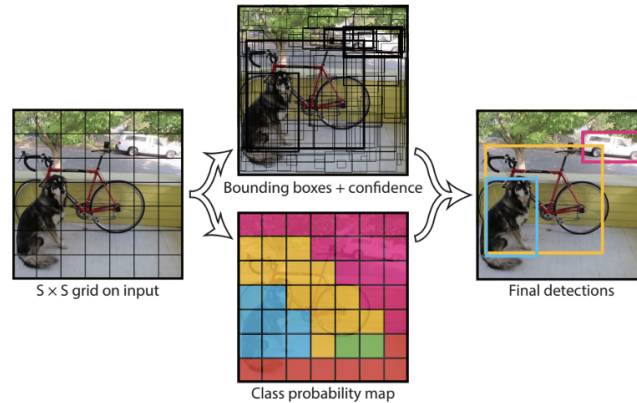


Figure 2.3: YOLO Approach

The model divides the image into an $S \times S$ grid, and for each cell, predicts a number of bounding boxes, the confidence for those boxes and the class probabilities.

2.1.3 Comments

As seen from the research, we can clearly see that there are many ways to solve the human detection problem. The classical approaches, although computationally efficient, are significantly outperformed by the deep learning approaches. For a mobile robot in a public area, we want to be able to detect almost all humans in the surroundings to better inform the PWU.

However, the major disadvantage of the deep learning approach is the time taken to train the network, as well as the requirement of a GPU to achieve real-time performance. These issues will be addressed in a later section of the report.

2.2 Object Tracking

2.3 Head Pose Estimation

2.4 Body Pose Estimation

2.5 SLAM

2.6 Augmented Reality Headsets

Chapter 3

Requirements Capture

Bibliography

- [1] Ya Li Hou and Grantham K.H. Pang. Human detection in crowded scenes. *Proceedings - International Conference on Image Processing, ICIP*, (September 2010):721–724, 2010.
- [2] Dongdong Zeng, Ming Zhu, Tongxue Zhou, Fang Xu, and Hang Yang. Robust Background Subtraction via the Local Similarity Statistical Descriptor. *Applied Sciences*, 7(10):989, 2017.
- [3] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. Technical report.
- [4] Massimo Piccardi. Background subtraction techniques: a review*. 2004.
- [5] Manato Hirabayashi, Shinpei Kato, Masato Edahiro, Kazuya Takeda, Taiki Kawano, and Seiichi Mita. GPU Implementations of Object Detection using HOG Features and Deformable Models. Technical report.
- [6] Paul Viola and Michael Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. Technical report, 2001.
- [7] Navneet Dalal, Bill Triggs, Navneet Dalal, and Bill Triggs. Histograms of Oriented Gradients for Human Detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.
- [8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.