Faculty of Engineering and Materials Science
German University in Cairo

Vision Based Motion Tracking System for Indoor Environments

A thesis submitted in partial fulfilment of the requirements for the degree of Bachelor of Science in Mechatronics Engineering

By

Anthony Rezkalla

Supervised by

Dr. Omar Shehata
Dr. Catherine Elias
Dr. Dalia Mamdouh
Dr. Abdelrahman Hatem

May 16, 2023

This is to certify that:

(i) the thesis comprises only my original work toward the Bachelor Degree of Science (B.Sc.) at the German University in Cairo (GUC),

(ii) due acknowledgment has been made in the text to all other material used

_____

Anthony Rezkalla
May 16, 2023

Acknowledgments

I would like to thank the following for their efforts with me in my study and for tolerating me during this journey...

Abstract

The abstract of the document is added here...
It is usually written after finishing writing all the other chapters.

Contents

List of Abbreviations

MRSs	Multi-Robot Systems
UAVs	Unmanned Aerial Vehicles
UGVs	Unmanned Ground Vehicles
HOG	Histogram of Oriented Gradients
SVM	Support Vector Machine
YOLO	You Only Look Once
Faster R-CNN	Faster Region-based Convolutional Neural Network
SSD	Single Shot Detection
RoIs	Region of Interests
RPN	Region Proposal Network
CNN	Convolution Neural Network

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Section Name

Some sample text with an Unmanned Aerial Vehicles (UAVs), some citation [1, 2], and some more Unmanned Ground Vehicles (UGVs).

1.2 Another Section

Reference to Section 1.1, and reuse of UGVs nad UAVs with also full use of Multi-Robot Systems (MRSs). Reference to figure 2.8.

Figure 1.1: GUC Logo

Chapter 2

Literature Review

Motion detection systems have a long history, with advancements in technology enabling the development of more sophisticated and accurate solutions over time. In the context of indoor environments, vision-based motion detection systems have played a crucial role in enhancing security, automation, and surveillance applications. Let's explore the historical background of these systems.

Numerous methods and approaches have been employed for the purpose of classifying and identifying individuals or objects. These techniques encompass various methods and serve the purpose of distinguishing and detecting people or objects. To begin our discussion, let's start by focusing on the initial technique known as background subtraction.

2.1 Background Subtraction

Background subtraction also called image difference algorithm is the initial method employed for detecting motion. A study conducted by Kalpesh R Jadav from GTU University [3]. presents a project aimed at developing a system for detecting and tracking moving objects using a stationary camera. The system's objective is to estimate parameters such as distance and velocity. The project uses a vision system that employs the image difference algorithm for general moving object detection and tracking.

The article emphasizes the detection of moving objects in a given scene, tracking of the detected objects, and estimation of their position and velocity. MATLAB software is used to implement the algorithm, and it enables the calculation of distance, frame per time, and velocity. Jadav's study also proposes an algorithm for estimating the velocity of a moving object using an image processing technique based on the camera calibration parameters and MATLAB software.

(a) Single object moving in background (b) Single object detected (c) Tracking of moving object

Figure 2.1: Image Difference Algorithm

Background subtraction is a technique used to detect objects in an image or video sequence by comparing each frame to a reference background image. The idea is that the reference background image represents the static background of the scene, while objects that move in the scene will cause changes in the pixels of the frames.

The background subtraction algorithm identifies the moving objects by subtracting the reference background image from each new frame in the sequence. The resulting difference image represents the areas where the pixels have changed in the new frame compared to the reference background image. If the difference image exceeds a certain threshold, it is considered as an object and detected by the algorithm as seen in figure 2.1.

Here is the general process of background subtraction:

1. Capture a reference image (background) of the scene without any objects or with only static background objects.
2. Obtain subsequent video frames and subtract the reference image from each of them.
3. Threshold the resulting difference image to identify the moving objects.
4. Perform noise reduction and image segmentation on the image to improve the detection accuracy.
5. Finally, track and analyze the detected objects to extract useful information about them.

Background subtraction can be used in various applications like surveillance and traffic monitoring. One of the major challenges of background subtraction is the need for accurate background modeling, since the system relies on the quality of the reference background image. This means that the algorithm can produce inaccurate results if there are significant changes in the scene's background or lighting condition. Thus, different techniques are used to improve the algorithm's performance, such as adaptive background modeling and temporal smoothing.

2.2 Histogram of Oriented Gradients (HOG)

Histogram of Oriented Gradients (HOG) is a feature extraction technique widely used in computer vision and image processing for object detection and recognition. The HOG feature descriptor works by capturing the distribution of edge directions and gradients in an image and representing them as a feature vector that can be used to identify objects in images.

Here is how HOG works:

1. Image pre-processing: The first step in HOG is to preprocess the image. This includes converting it to gray scale, applying Gaussian smoothing, and calculating the gradients in both the x and y directions.
2. Divide image into cells: HOG divides the image into small fixed-size cells, typically 8x8 or 16x16 pixels.
3. Orientation binning: Within each cell, the gradient orientations of each pixel are accumulated into orientation bins. Typically, these bins are 9 or 10 bins evenly spaced from 0 to 180 degrees.
4. Normalization: Before the feature vector is calculated, each block (a group of cells) undergoes normalization to make the HOG descriptor less sensitive to changes in lighting and contrast.
5. Feature vector: Finally, the feature vector is calculated by concatenating the normalized histograms of the cells within each block.

The resulting feature vector is often used as input to a Support Vector Machine (SVM) classifier or other machine learning algorithm for object detection and recognition tasks and the output is shown as figure 2.2.

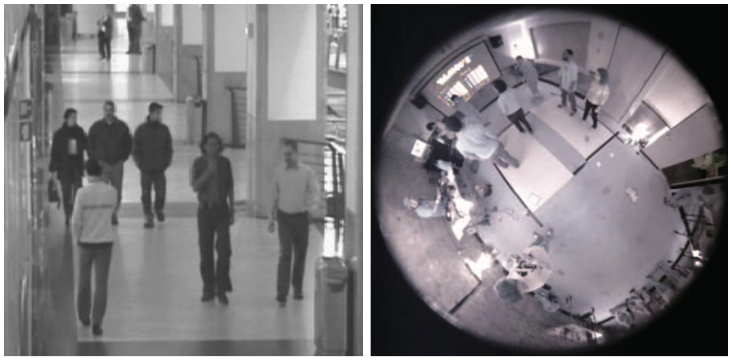
(a) Input Image (b) Output Image

Figure 2.2: Histogram of Oriented Gradients

Fish-eye cameras can provide a 360-degree view of a large area using a single camera, making them efficient. However, detecting humans in fish-eye images remains a challenge. To address this, a human detection algorithm based on Histogram of Oriented Gradient (HOG) features is proposed by An-Ti Chiang from New York University [4].

This algorithm involves rotating each search window on a radial line to the vertical reference line and training a SVM classifier using HOG with positive and negative examples obtained through such rotations as shown in Figure 2.8.

To detect humans in an image, the image is rotated successively and windows containing humans along the reference line are detected using the trained classifier. Multiple window sizes are used to detect people of different sizes and an algorithm is developed to find overlapping windows covering the same person and identify the best enclosing window. This method has yielded highly accurate human detection in low-resolution and low-contrast images with multiple people of different poses and sizes as shown in 2.4.



(a) (b)

Figure 2.3: (a) conventional camera view (b) fish-eye camera view.

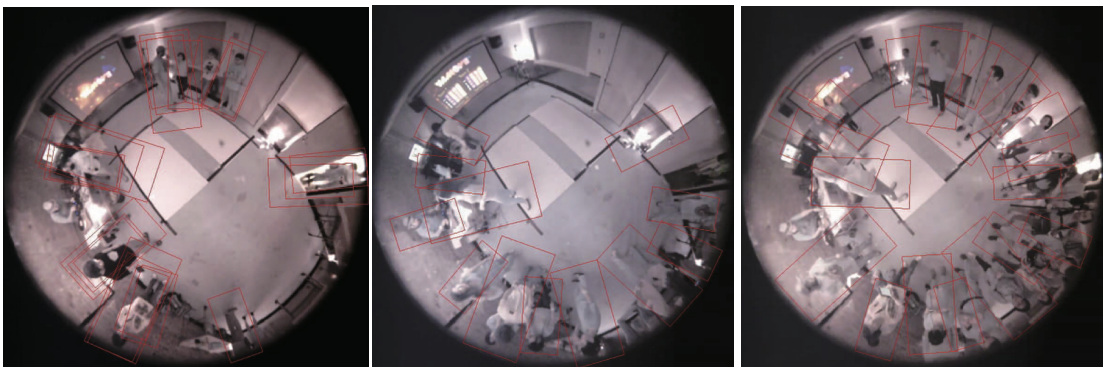


Figure 2.4: HOG results with Fish-Eye view

In summary, the HOG feature descriptor is a powerful technique for object detection and recognition in computer vision and image processing. It works by capturing the distribution of edge directions and gradients in images and represents them as a feature vector for further analysis.

2.3 Haar Cascade Classifiers

Haar Cascade Classifiers are a popular computer vision algorithm used for object detection. It was named "Haar" because it uses a mathematical function called Haar Wavelets.

The algorithm works by training a model on a large data-set containing samples of both positive and negative images. A positive image contains the object that the algorithm is being taught to recognize, while negative images do not. The classifier then uses the extracted features from these images to learn how to differentiate between images containing the object and those that do not.

The extracted features are based on Haar wavelets, which are small, square-like patterns as showin in figure**??**that can be used to detect edges and textures in an image. The classifier combines these features into a sort of "template" that matches the object being searched for.

Once the Haar Cascade Classifier has been trained on the data-set, it can be used to detect the object in new images. The classifier slides this "template" over the new image, searching for matches. If a match is found, it is marked as a positive detection. Haar Cascade Classifiers are known for their accuracy, speed, and low computational requirements.

Therefore, Haar Cascade Classifiers can detect various objects in real-time, including faces, pedestrians, vehicles, and even more complex objects. It's used widely in various tech areas such as security, automotive advanced driver assistance systems, and robotics.

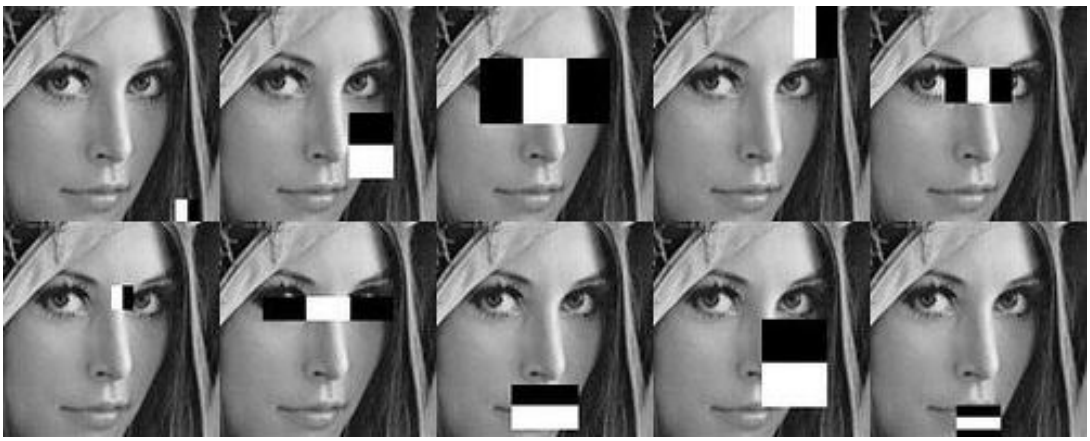


Figure 2.5: Haar Cascade Classifier Example

A project by Ahmad Adlan from Narotama University [5] investigates the use of facial recognition for a surveillance system. The typical video surveillance system employs closed-circuit television (CCTV) technology to record video for security purposes and usually requires manual identification of individuals through their appearance on recorded video. However, modern surveillance camera systems generally do not include facial recognition capabilities.

The proposed system utilizes a surveillance camera system with the Haar cascade classifier to automatically identify the identity of individuals using facial recognition in real-time. The hardware used in the project features Raspberry Pi as a processor and Pi Camera as a camera module. The project was developed in three phases, which involve data gathering, training recognizer, and face recognition using Python programming and the OpenCV library as in Figure 2.6.

The system successfully displays the output of human facial recognition with facial angles within ±40°, under medium and normal light conditions, and at distances between 0.4 and 1.2 meters. Targeted images are allowed to wear face accessories as long as the accessory does not cover the facial structure.

This system can significantly reduce the cost of manpower and streamline the process of identifying individuals in real-time surveillance situations, potentially providing a useful tool for security purposes.

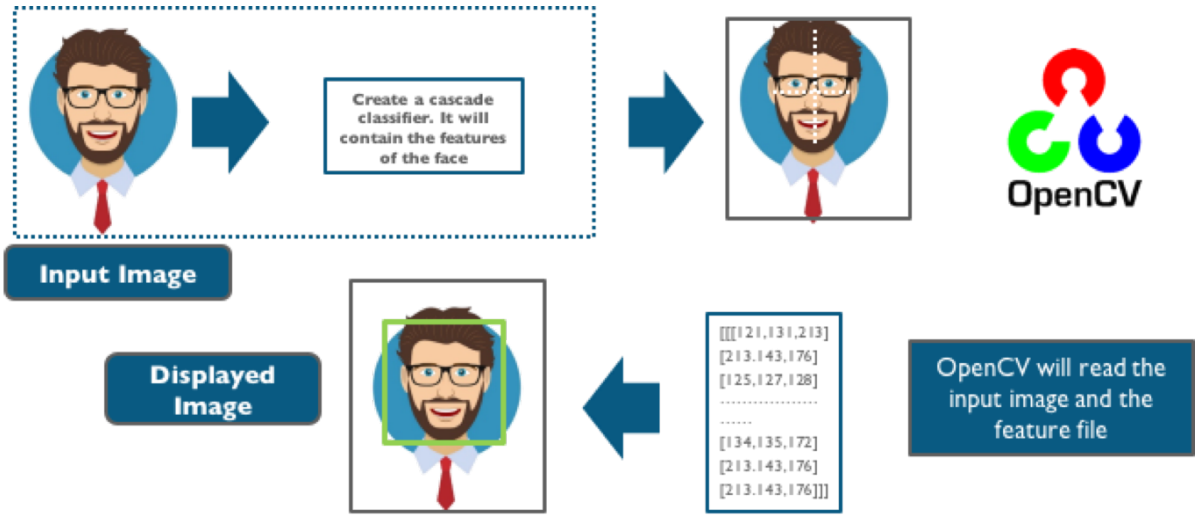


Figure 2.6: Haar Classification

2.4 Deep Neural Networks

Deep neural networks are also known as deep learning, a subset of machine learning that is based on the structure of the human brain. It consists of artificial neural networks with more than two hidden layers.

Deep neural networks are designed to learn by adjusting the connection weights between individual neurons through back-propagation. During the training phase, the network receives input data, propagates it forward through the network, and adjusts the weights to minimize the difference between the predicted output and the known target output. This process continues until the network produces accurate predictions on unseen data. As shown in figure 2.8 the layers between the input and output layers are called hidden layers, where the data transformation and feature extraction occur.

Deep neural networks are used for a variety of applications such as image and speech recognition, natural language processing, anomaly detection, and recommendation systems. They have shown remarkable performance in the fields such as computer vision, robotics and machine translation.

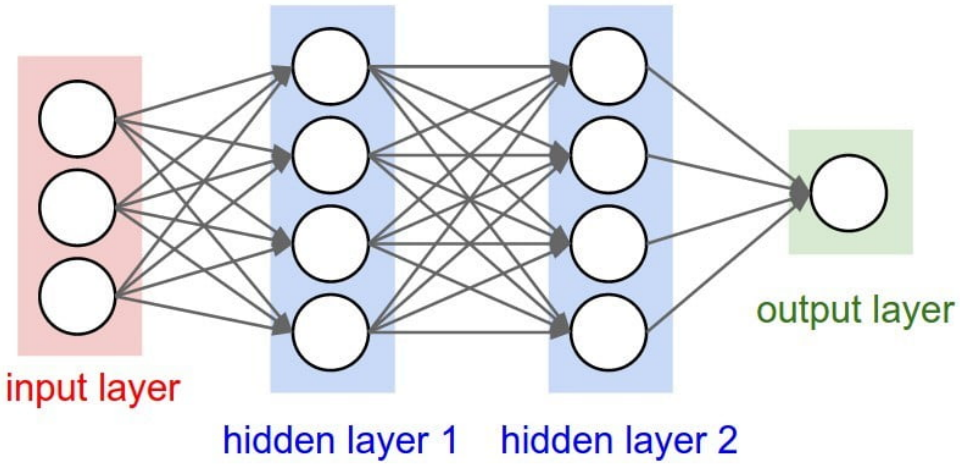


Figure 2.7: Deep Neural Network

Here are 3 common ways people used Deep Neural Networks to implement vision object detection:

2.4.1 You Only Look Once (YOLO)

You Only Look Once (YOLO) is a real-time object detection system that uses deep neural networks. It is one of the most popular and efficient object detection algorithms used to label objects within images or real-time video.

YOLO works by dividing an image into a grid, and for each cell of that grid, the model predicts bounding boxes, confidence scores and class probabilities. The confidence score represents how confident the model is about an object being present in a grid cell. Each bounding box consists of four values that represent the predicted coordinates of the object's top-left corner, width, and height. The class probabilities are associated with each bounding box and represent the probability of the presence of a particular object class within that bounding box.

The YOLO architecture is based on a custom deep neural network that divides an image into a grid and then applies a set of convolutional filters to this grid. The feature maps captured by these filters are then used to identify the potential bounding boxes. The bounding boxes are then refined using another set of convolutional filters and non-maximum suppression is applied to remove redundant bounding boxes.

One of the major advantages of YOLO is that it is very fast, capable of processing up to 45 frames per second on a high-end GPU. YOLO also performs well in cluttered scenes, and it can detect objects at various angles, scales, and orientations. Additionally, the YOLO architecture allows for end-to-end training, which means that the whole network can be trained using a single loss function.

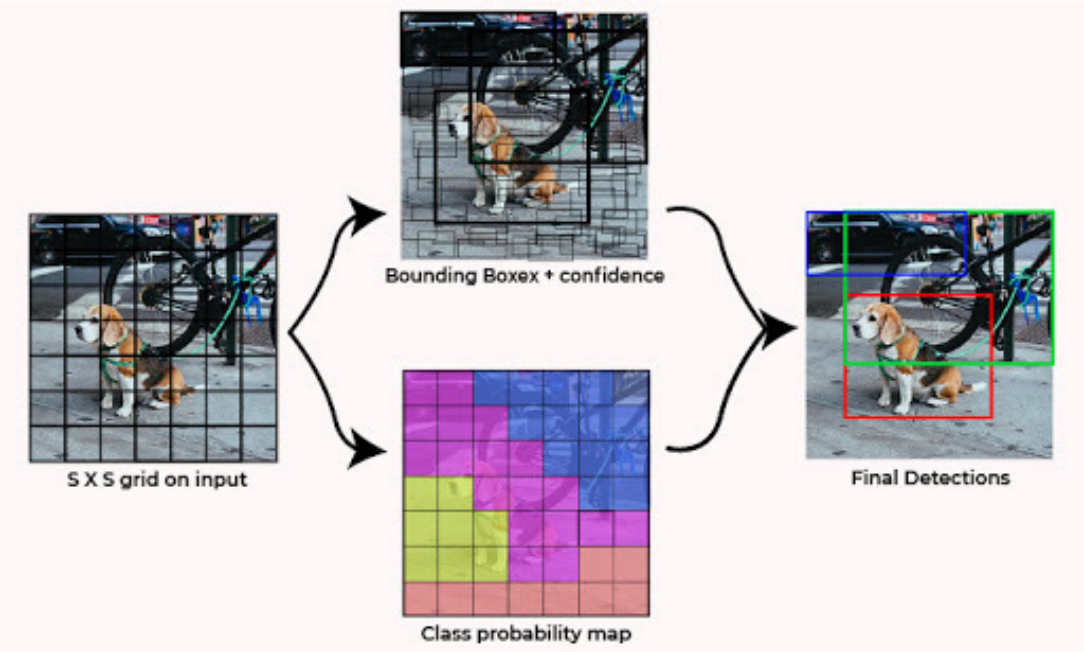


Figure 2.8: YOLO Technique

A project by Kiran Jot Singh from University of Eastern Finland [6] presents a viable approach for real-time computer vision based object detection and recognition to aid a mobile robot in efficient indoor navigation. The future is foreseen as a collaboration of humans with mobile robots to assist in everyday tasks.

Mobile robotic systems are widely used for home assistance, emergency services, and surveillance, where real-time critical actions need to be taken within seconds. The proposed algorithm is an enhancement of the YOLO algorithm that offers improved object detection and recognition capabilities as shown in Figure 2.9 while requiring fewer computational resources and having a smaller network structure.

The algorithm's effectiveness was compared to other conventional object detection/recognition algorithms based on mean average precision (mAP) score, inference time, weight size, and false positive percentage. The study provides evidence that the proposed algorithm outperforms conventional methods.

The proposed framework utilizes the results of efficient object detection/recognition to assist the mobile robot's indoor navigation. The framework is expected to be useful in various indoor navigation robots for different services.

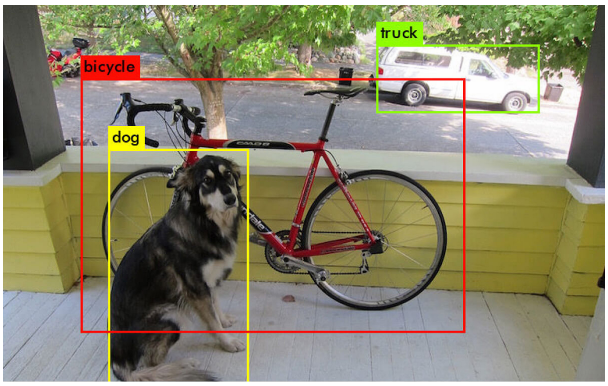


Figure 2.9: YOLO Technique Result

In summary, YOLO is a fast and effective object detection system that can detect objects in real time with high accuracy. It works by dividing an image into a grid, predicting bounding boxes, confidence scores, and class probabilities for each grid cell, which makes it an ideal solution for real-time object detection applications such as autonomous driving, drones and robotics.

2.4.2 Single Shot Detection (SSD)

Single Shot Detection (SSD) is a real-time object detection algorithm that uses a deep neural network for object recognition in images. It is a one-stage detector, which means that it performs the tasks of both object localization and classification in a single feed forward pass of a neural network.

SSD works by dividing an incoming image into a grid of default boxes. Each default box has a set of predicted offsets for the bounding box coordinates and a predicted class probability for each object category. The default boxes also have a set scale and aspect ratio, which helps the model adjust the bounding boxes to better fit the objects in the image.

SSD uses three different types of feature maps to capture object information at various scales. The first set of feature maps comes directly from the base network, and the other two sets of feature maps are generated by additional convolutional layers.

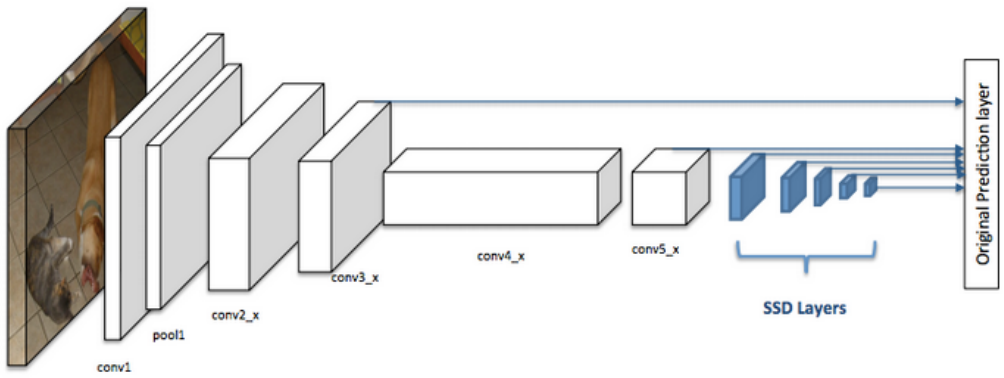


Figure 2.10: SSD Method

One of the major advantages of SSD is that it is fast and can process up to 59 frames per second on a GPU. However, it may not perform as well as two-stage object detection algorithms, such as Faster R-CNN mentioned next, since it has to depend on the grid of default boxes to identify objects. Nevertheless, SSD is highly effective, particularly in detecting small objects in densely populated areas, where other object detection algorithms have difficulty.

A study done by by Zihao Wang from University of Electronic Science and Technology of China [7] aimed to design benchmarks for object tracking with motion parameters (OTMP). A Fast Depth-Assisted Single Shot MultiBox Detector (FDA-SSD) algorithm is proposed, combining depth information into Single-Shot MultiBox Detector (SSD) to track 3D targets.

The algorithm effectively combines the 2D detection model and the 3D positioning algorithm, and a framework is established using monocular motion platforms for target detection and tracking. The detection model adapts to the spatial geometric constraints of the target to solve the target depth information. The normalized depth information is then utilized to select the feature window for the detector, significantly reducing computational requirements. Compared to the original SSD method, the network model has fewer operating parameters, while maintaining a high recognition rate consistent with the original SSD.

Experiments were conducted on target detection and tracking based on monocular motion platforms indoors. The spatial tracking trajectory's root mean square error (RMSE) was less than 4.72 cm, affirming the framework's effectiveness in achieving visual detection, classification, and spatial tracking using a monocular motion platform.

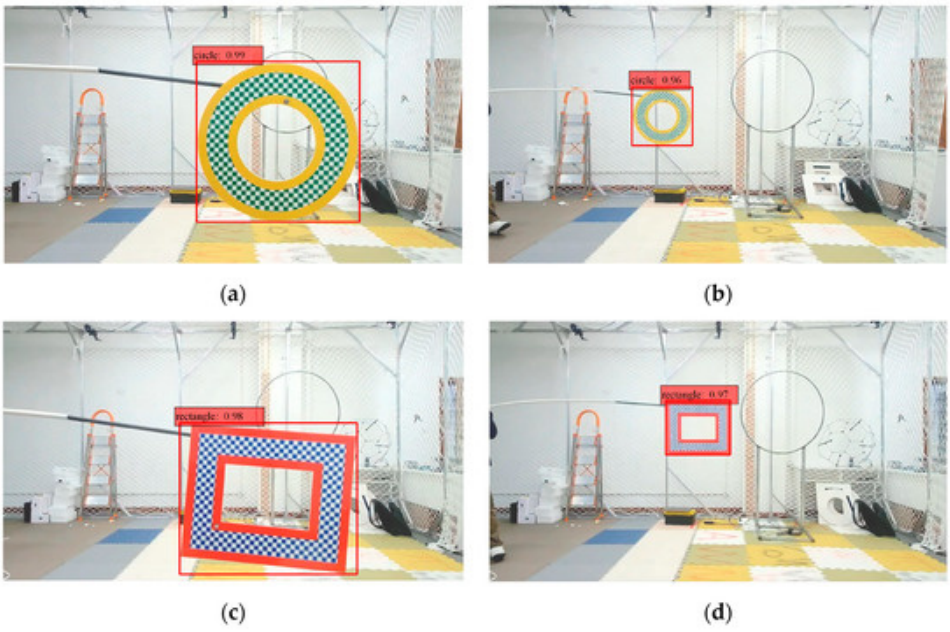


Figure 2.11: SSD Results

In summary, SSD is a real-time object detection algorithm that enables object classification and localization in a single feed forward pass of a neural network. It uses a set of default boxes to predict object categories and bounding box coordinates, making it effective at identifying small objects in densely populated areas. The algorithm is fast and well-suited for real-time applications like autonomous driving and robotics.

2.4.3 Faster Region-based Convolutional Neural Network (Faster R-CNN)

Faster Region-based Convolutional Neural Network (Faster R-CNN) is a two-stage object detection algorithm that uses deep neural networks to detect objects in images. It was developed by researchers at Microsoft and is known for its speed and accuracy.

The first stage of Faster R-CNN proposes a set of candidate object regions called Region of Interests (RoIs) using a Region Proposal Network (RPN). The RPN is a fully convolutional neural network that predicts bounding boxes and objectness scores for each position in the feature map. These RoIs are then passed to the second stage, which performs classification and bounding-box regression to localize and identify the objects.

In the second stage, the RoIs are pooled into a fixed size and are classified into object categories (e.g. car, person, dog) and regress bounding boxes around the object. The output is a set of proposed object detection along with a confidence score for each object category.

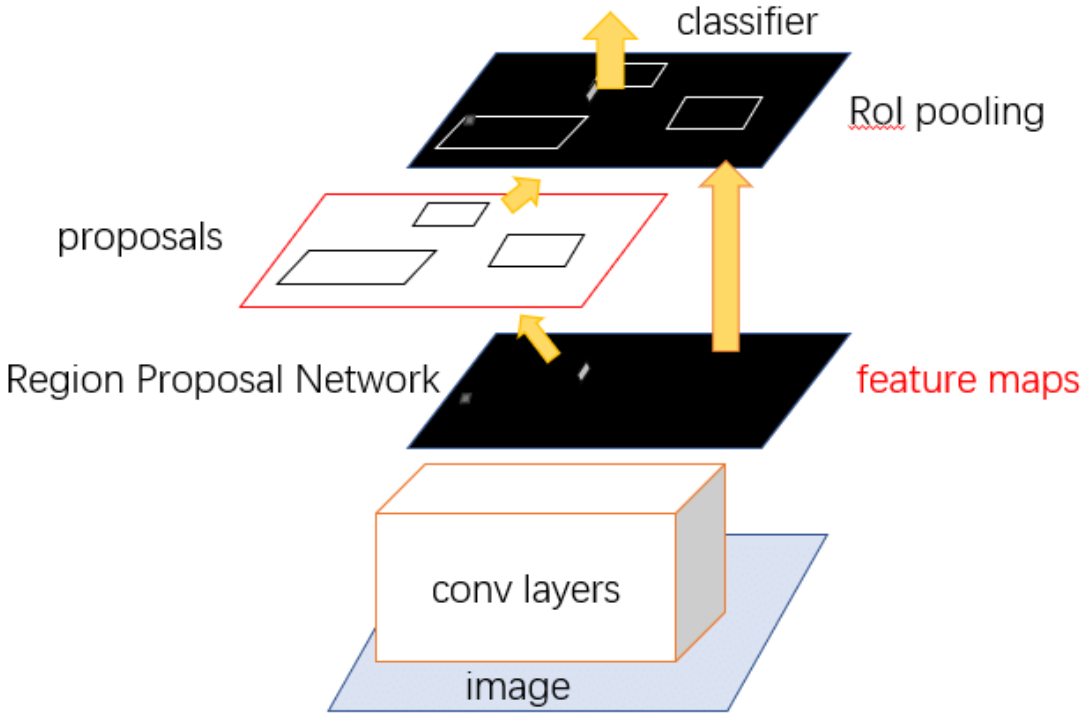


Figure 2.12: Faster R-CNN Method

One of the major advantages of Faster R-CNN is that it can detect objects with higher accuracy than other object detection algorithms, such as YOLO or SSD. It is also efficient compared to other two-stage object detection algorithms due to the region proposal network's shared computation with the object detection network, allowing it to process up to 7 frames per second.

A project by Wnedy Cai from The Library of Wuhan University of Technology [8] presented a system for detecting street objects. The Generic Model detection algorithm is based on Convolution Neural Network (CNN) requires the creation of a training model, which can be time-consuming during both the training and testing phases. To address this issue, Transfer Learning is used to fine-tune pre-trained models by leveraging the COCO image data-sets for specific deep learning models with customized weights and outputs.

Furthermore, the CNN structure is adjusted to improve overall performance. The training is then tailored to the street environment. After conducting experiments, the resulting fine-tuned found in figure 2.13 network has been found to be highly effective compared to traditional models.

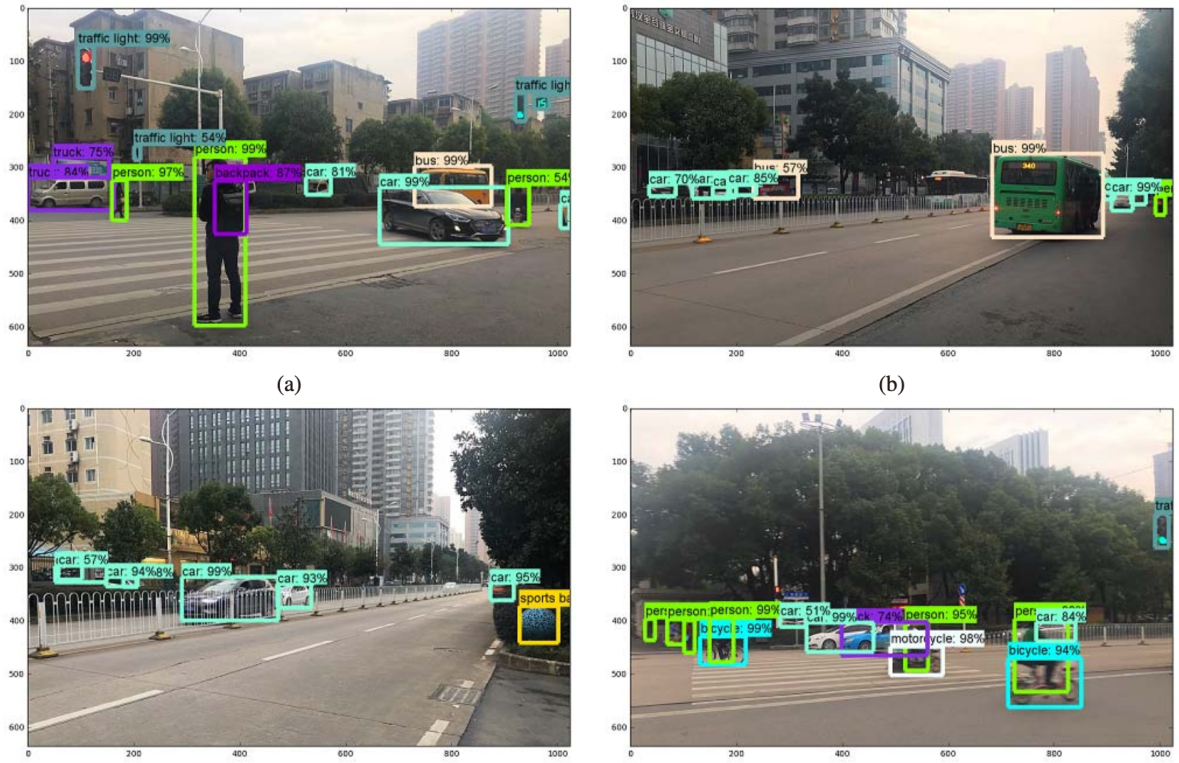


Figure 2.13: Faster R-CNN Results

Faster R-CNN is widely used for a wide range of applications, such as object detection in autonomous vehicles, robotics and computer vision. It has become one of the go-to methods for object detection tasks, and its widespread use is tied to the excellent accuracy-performance trade-off.

Chapter 3

Methodology

Equation (3.1) describes the kinematics model of the unicycle

$$\dot{q}_i = \begin{bmatrix} \dot{x}_i \\ \dot{y}_i \\ \dot{\varphi}_i \end{bmatrix} = S_1(q_i)u_i \quad , \quad u_i = \begin{bmatrix} v_i \\ \omega_i \end{bmatrix} \quad and \quad S_1(q_i) = \begin{bmatrix} \cos\varphi_i & 0 \\ \sin\varphi_i & 0 \\ 0 & 1 \end{bmatrix} \tag{3.1}$$

Chapter 4

Results

Chapter 5

Conclusion

Hi

Chapter 6

Future Work

Appendix

Bibliography

[1] W.G. Campbell. *Form and style in thesis writing*. Houghton Mifflin, 1954.

[2] S. Wenkang. An analysis of the current state of English majors' BA thesis writing [J]. *Foreign Language World*, 3, 2004.

[3] Ashish Kumar SAhu and Abha Choubey. Motion detection surveillance system using background subtraction algorithm. *International Journal of Advance Research in Computer Science and Management Studies*, 1(6), 2013.

[4] An-Ti Chiang and Yao Wang. Human detection in fish-eye images using hog-based detectors over rotated windows. 09 2014.

[5] Adlan Ahmad, Sharifah Saon, Abd Mahamad, Cahyo Darujati, Sri Mudjanarko, Supeno Nugroho, and Mochamad Hariadi. Real time face recognition of video surveillance system using haar cascade classifier. *Indonesian Journal of Electrical Engineering and Computer Science*, 21:1389, 03 2021.

[6] Kiran Jot Singh, Divneet Kapoor, Khushal Thakur, Anshul Sharma, and Xiao-Zhi Gao. Computer-vision based object detection and recognition for service robot in indoor environment. *Computers, Materials Continua*, 72:197–213, 01 2022.

[7] Zihao Wang, Sen Yang, Mengji Shi, and Kaiyu Qin. Fda-ssd: Fast depth-assisted single-shot multibox detector for 3d tracking based on monocular vision. *Applied Sciences*, 12(3):1164, 2022.

[8] Wendi Cai, Jiadie Li, Xie Zhongzhao, Tao Zhao, and Kang LU. Street object detection based on faster r-cnn. pages 9500–9503, 07 2018.