IMPLEMENTATION OF A CLOUD BASED EMBEDDED PLATFORM FOR OBJECT DETECTION AND RECOGNITION

by

OLALEKAN ZAINAB TEMITOPE
(1832322648)

A dissertation submitted in partial fulfillment of the requirements for the degree of
Master of Science (Embedded System Design Engineering)

School of Computer and Communication Engineering
UNIVERSITI MALAYSIA PERLIS

2019

UNIVERSITI MALAYSIA PERLIS

DECLARATION OF THESIS

Author's Full Name	:	OLALEKAN ZAINAB TEMITOPE
Title	:	IMPLEMENTATION OF A CLOUD BASED EMBEDDED PLATFORM FOR OBJECT DETECTION AND RECOGNITION
Date of Birth	:	15 OCTOBER 1991
Academic Session	:	2018/2019

I hereby declare that this dissertation becomes the property of Universiti Malaysia Perlis (UniMAP) and to be placed at the library of UniMAP. This dissertation is classified as:

□	**CONFIDENTIAL**	(Contains confidential information under the Official Secret Act 1997)*
□	**RESTRICTED**	(Contains restricted information as specified by the organization where research was done)*
☑	**OPEN ACCESS**	I agree that my dissertation to be published as online open access (Full Text)

I, the author, give permission to reproduce this dissertation in whole or in part for the purpose of research or academic exchange only (except during the period of _______ years, if so requested above)

Certified by:

SIGNATURE

SIGNATURE OF SUPERVISOR

A07742957

(NEW IC NO. /PASSPORT NO.)

Date:

Dr Said Amirul Anwar Ab. Hamid @ Ab Majid

NAME OF SUPERVISOR

Date:

NOTES : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with the period and reasons for confidentiality or restriction.

ACKNOWLEDGMENT

I will like to express my profound gratitude to Allah, who guided me in all stages and enabled me to complete this thesis. This work would not have been completed without the aid of a lot of people. Firstly, my parents Mr. and Mrs. Olalekan who sacrificed immensely (both financially, emotionally) to my study in Malaysia. You have always believed in me, taught and encouraged me to spread my wings and soar. A special gratitude goes to my sister Comfort Kolapo who contributed to my study, family and, friends for the support. cheer and words of advice.

I would like to extend my sincere gratitude to my supervisor Dr. Said Amirul Anwar for his expertise, patience, encouragement, aid and guidance to this project work. I am grateful to all the lecturers who impacted knowledge in me during the course of this program.

I would also like to extend my sincere appreciation to all the staff of the School of Computer and Communication Engineering for the care and support throughout the course of this program.

I would not forget the members of the Nigeria community UNIMAP who took me as a friend and a sister. May God bless you, strengthen you and grant you excellence in all your endeavors.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AP	Average Precision
API	Application Programming Interface
CSI	Camera Serial Interface
CNN	Convolutional Neural Network
DSP	Digital Signal Processing
FC	Fully-Connected
FGPA	Field Programmable Gate Array
GPU	Graphics Processing Unit
IOU	Intersection over Union
mAP	mean Average-Precision
NMS	Non-Maximum Suppression
R-CNN	Region-based Convolutional Neural Network
RGB	Red, Green, Blue
SBC	Single Board Computer
SIFT	Scale Invariant Feature Transform
SSH	Secure Shell
SURF	Speeded Up Robust Features
VNC	Virtual Network Computing

LIST OF SYMBOLS

$\Delta_{\text{x-center}}$	Change in center of x
$\Delta y_{\text{-center}}$	Change in center of y
Δ_{width}	Change in center of x
Δ_{height}	Change in center of x
s	Seconds

Pelaksanaan Platform Terbenam Berbasis Awan untuk Pengesanan Objek dan Pengiktirafan

ABSTRAK

Dengan kemajuan baru-baru ini dalam model visi komputer berasaskan pembelajaran yang mendalam, pengesanan objek dan aplikasi pengiktirafan seperti pengawasan video, Bio-Imaging, kereta autonomi semakin meningkat. Teknik pengesanan objek memerlukan beberapa dataset imej, memori, mesin dengan GPU untuk melatih algoritma dan mempunyai penggunaan kuasa yang tinggi. Platform terbenam dicirikan oleh penggunaan kuasa rendah, ruang, dan sumber tenaga yang membuat penggunaan algoritma pada mereka sukar. Untuk mengatasi kelemahan ini, algoritma pengesanan (Faster R-CNN) dilatih dan diuji dengan dataset imej yang diperoleh dari ImageNet. Algoritma ini dilaksanakan pada komputer dengan MATLAB. Peranti pengambilan gambar dibentuk menggunakan Raspberry pi dan pi kamera untuk menangkap, memproses dan menghantar imej ke pengesan melalui platform cloud Dropbox dengan Python. Platform Dropbox berfungsi sebagai antara muka antara Raspberry pi dan pengesan jauh. Pengesan dilatih untuk mencari lima kelas objek iaitu Broom, Fan, Keyboard, Mouse, dan Televisyen. Pengesan objek berbilang kelas telah dilatih pada 2500 imej dengan setiap kelas mempunyai 500 imej pegun dan diuji pada 500 imej pegun. Sistem ini diuji dalam masa nyata dengan menangkap imej pada Raspberry pi dan menghantarnya ke dan dari pengesan menggunakan akses internet untuk menentukan tempoh proses. Ketepatan pengesan diukur menggunakan metrik ketepatan purata (AP) untuk setiap kelas dan mengira metrik purata ketepatan purata (mAP) untuk semua kelas. Pengesan objek berbilang kelas mencapai Purata Purata Ketepatan (mAP) sebanyak 0.67 dan keseluruhan prosedur sistem dari imej yang ditangkap ke paparan akhir dilaksanakan dalam purata 45 saat.

Implementation of a Cloud-Based Embedded Platform for Object Detection and Recognition

ABSTRACT

With the recent advancements in deep learning-based computer vision models, object detection and recognition applications such as video surveillance, Bio-Imaging, autonomous cars are increasing in number. Object detection techniques require some large image datasets, memory, a machine with GPU to train the algorithm and have high power consumption. Embedded platforms are characterized by low power consumption, space, and energy resources making the deployment of the algorithms on them difficult. In order to overcome these drawbacks, the detection algorithm (Faster R-CNN) is trained and tested with an image dataset obtained from ImageNet. This algorithm is implemented on a computer with MATLAB. An image acquisition device is set up using the Raspberry pi and pi camera to capture, process and send images to the detector via Dropbox cloud platform with Python. The Dropbox platform serves as an interface between the Raspberry pi and the remote detector. The detector was trained to locate five classes of objects which namely Broom, Fan, Keyboard, Mouse, and Television. The multi-class object detector was trained on 2500 images with each class having 500 still images and tested on 500 still images. The system was tested in real-time by capturing images on the Raspberry pi and transmitting it to and from the detector using internet access in order to determine the process duration. The detector accuracy is measured using the average precision (AP) metric for each class and calculating the mean average precision (mAP) metric for all classes. The multi-class object detector achieved a mean Average Precision (mAP) of 0.67 and the entire system procedure from image capturing to the final display was executed in an average of 45 seconds.

CHAPTER 1 : INTRODUCTION

1.1 Overview

In recent years, computer vision domain has not been left out from thriving due to the endless efforts of researchers and significant advancement in related fields. Computer vision uses digital images to model and emulates human vision using a computer through three major steps. These steps are image acquisition, image processing with the third being image analysis and understanding. As a result of this, applications of computer vision such as pattern recognition, medical imaging, 3D model building, surveillance, object detection and recognition, face detection has been made a reality.

Object detection is basically determining instances of real-world objects in images or videos while object recognition is the identification of target objects in still images or videos (MathWorks, 2019). Feature extraction is the first step when detecting an object in an image. It is a process whereby key points in an image is located and obtaining the required information in that location. Good Features to Track (GFTT) and Scale Invariant Feature Transform (SIFT) algorithms were used to extract important features of an object in images. These algorithms, however, involved heavy floating points calculations and were computationally complex making them unsuitable for real-time embedded platforms. The Speeded Up Robust Features (SURF) algorithm was proposed by Bay et al. in ECCV 2006 conference. Its performance is on par with SIFT and is faster compared to the former (Panchal et al., 2013). The major advantage of SURF is that it uses an integral image for feature detection and description which heightens the process efficiency. Nevertheless, SURF like other feature-based algorithms is computationally expensive and frequently brings about a very low frame rate (Zhao et al., 2013).

The emergence of intelligent features in notes, tablets, surveillance, smartphones and automotive systems which have ready-made features for capturing images lead to the need for more advanced algorithms with lesser computational complexity and higher accuracy. This resulted in the introduction of a deep learning algorithm which is vector accelerated like Convolutional Neural Networks (CNN) for classification, localization, and detection of objects in images. As regards deep learning, object detection is a subset of object recognition, where the target object is not only identified but also located in still images and videos. As depicted in Figure 1.1, bounding boxes are drawn around the objects to be identified which in this case is a person and a dog. The CNN is a deep learning algorithm that is flexible, adaptive and higher in its accuracy. It allows quick tuning to new objects without changing the algorithm. The level of accuracy of the CNN algorithm is dependent on the quantity of the image dataset. The algorithm yields a less accurate for smaller data but shows significant accuracy on the large image datasets. Hence, CNNs require a large number of labelled datasets to perform computer vision-related tasks (Pathak, Pandey, & Rautaray, 2018).

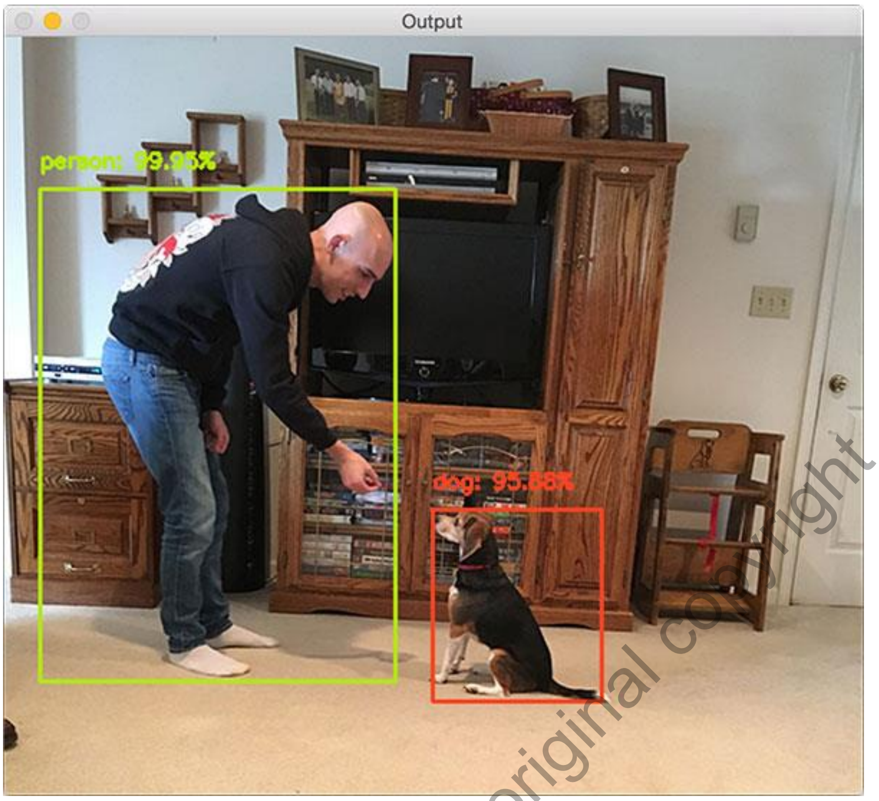

Figure 1.1 Localization and Identification of objects
Source: Pyimagesearch (2017)

1.2 Problem Statement

Traditionally, image acquisition devices are bulky, complex and expensive. With the growth of the various field in technology, prices and sizes of these systems have reduced significantly. Image acquisition is one of the main processes in computer vision where an image is captured and sent to a computer or embedded device through an interface. It is an important preliminary action taken to generate image data (Martynenko, 2017). Conventional image acquisition systems have slow processing speed and require a computer or workstation to pre-process the images. Furthermore, due to their bulkiness, they are not portable and are difficult to be integrated with other systems. In an efficient detection system, it is pertinent that tasks of capturing, displaying as well as the processing of images are incorporated (Liu, Liang, & Cheng, 2011). The availability of embedded platforms which are reliable, portable, easy to use, have low energy consumption with prices on the low side has increased (Fuente, González-Castro, Fernández-Robles, & Alegre, 2015). These merits make the embedded system suitable for the role of acquiring images.

The emergence of deep convolutional neural networks largely boosted the development of several artificial intelligence applications. These networks are often filled with hundreds to thousands of interconnected layers in which computation of millions of parameters from a frame of sensor data is required for a single classification (Lane et al., 2017). Deep neural networks such as CNN performs excellently compared to traditional algorithms at the cost of complex computation and energy consumption. Fast Region-Based Convolutional Neural Network (Fast R-CNN) is a deep neural network introduced by Girshick, (2015a) as a further development to the previous work which is Region-Based Convolutional Neural Network (R-CNN) to detect and classify objects in images. Lee, Son, Kim, & Park, (2017) stated that general-purpose computer with the only CPU becomes heavily loaded and more often cannot achieve the real-time requirements when implementing the algorithm except when equipped with Graphics Processing Units (GPU). To overcome the limited computing resources and energy supply of the embedded systems, Mao et al., (2018) implemented Fast R-CNN on the board by modifying the algorithm to fit the specific embedded platform or implementing a CPU + GPU platform. Although this solution gave a satisfactory result, the cost and complexity of the implementation increased.

Significant resources have been expended towards building user-friendly and smart applications on embedded devices. The application of deep neural network on these devices could amount to a generation of applications that is capable of performing recognition tasks to support a higher level of interaction between man and his physical environment (Yao et al., 2018). In the Internet of Things (IoT) hardware design, embedded platforms such as Raspberry pi with minimal specifications such as little computational capability (1Gigabyte of RAM), energy resources (works on 5V and limited memory) compared to computer or workstation are typically used. These IoT devices can be designed for any application and able to transmit and receive data through the network. The Raspberry pi is a Single Board Computer (SBC) that is affordable, small-sized and is easy to work with, but with the limited resources needed to implement the deep learning algorithms. In addition, a significant amount of memory is required to store the large dataset used when training the deep neural network. The need to evaluate the performance of the proposed IoT system arises for further improvement.

1.3 Research Questions

The goal of this research is to design an embedded system which detects and recognize objects in images. The first question that comes to mind is if the SBC can execute the detection algorithm which in this case is Faster R-CNN. The next question is if the Raspberry pi has good and stable connectivity to transmit and receive data. Moreover, if the object detection and recognition algorithm can be executed on a remote computer serving in a short amount of time. Lastly, whether the developed system gives a good performance.

1.4 Research Objectives

The objectives of this research are summarized as follows:

1. To develop an image acquisition and processing platform using Single Board Computer (SBC).

2. To implement Faster R-CNN algorithm for object detection and recognition.

3. To evaluate the overall performance of the system in terms of accuracy and processing time.

1.5 Research Scope

In the proposed study, the Raspberry pi 3 model B and Python programming language is used to develop the image acquisition device are used while MATLAB is used for the object detection and recognition algorithm. The Faster R-CNN algorithm will be evaluated by its accuracy in detecting and recognizing the objects in the image. This work is limited to RGB images downloaded from ImageNet containing five object classes (Broom, Fan, Keyboard, Television, Mouse) for training and testing the algorithm. Real scene images which were captured by the Raspberry pi is used to evaluate the whole system. The cloud storage platform used as an interface in this work is Dropbox.

1.5.1 Dissertation Organization

This dissertation is organized into five chapters and the composition of each chapter as summarized as follows:

1. Chapter 1 presents the background of the subject matter as well as the problem statement, objectives and scope of the research. The layout of the dissertation is also included.

2. Chapter 2 is the literature review which consists of the comparison between some Neural Networks as well as a review of the past study on deep learning techniques. Previous works of image processing using the FPGA, DSP, and SBC is revised and lastly, general information and background of the Raspberry pi are discussed.

3. Chapter 3 involves the methodology. In this section, the description of the technique applied in the study, procedures, the block diagram of the work as well as the Raspberry pi setup is explained.

4. Chapter 4 portrays the results of the object detector training and testing on the ImageNet dataset offline. Likewise, the outcome of the system when an image is captured by the acquisition device and processed online with a detailed explanation of it.

5. Chapter 5 concludes the achievements of this study objectives with future work recommendation.

CHAPTER 2 : LITERATURE REVIEW

2.1 Introduction

In recent decades, the domain of computer vision has witnessed researches due to the vast applicability in fields such as video surveillance, robotics, autonomous systems, and scene understanding. Object detection and object recognition are related techniques that play vital roles in the computer vision domain. According to Pathak et al. (2018), object detection is carried out by determining the instance of the class which the object belongs and putting a bounding box around the object in order to estimate its location. Detection of an object can be a single class object detection where a single instance of the class is detected from an image or multi-class object detection in which the classes of all object in the image is detected (Pathak et al., 2018). Generally, the object detection task is in three-phase, candidate regions are selected, features are extracted based on these regions, after which the classification task is done using pre-trained models (Guan & Zhu, 2017). Object recognition aims at the accurate identification of the target object from an image (Wu, Bie, Guo, Meng, & Zhang, 2017). it involves identifying a target object in an image from a series of well-known tags. The techniques of object recognition can be classified based on 2D or 3D image information. Face, leaf, fruit, pattern, alphabets are examples of objects that can be detected and recognized.

In order to detect an object, an idea of the possible position of the object and how the image is segmented is needed. This poses a kind of chicken-and-egg problem whereby the location of an object should be known so as to recognize its shape and class, and to know its location, the shape of the object is needed(Dirk, Laurent, Maximilian, Tomaso,

& Christof, 2002). Visually dissimilar features like clothes, shoes, bag, face or a person may be part of a particular object but for the detector to know this, the object must be recognized first while some objects slightly stand out from the background thus requiring separation before it is recognized(Uijlings, Sande, Gevers, & Smeulders, 2013).

According to Kumaran & Reddy (2017), object detection can be generally classified into three; motion-based detection, feature-based detection, and template-based detection. In the motion-based method, the detection depends on the grouping of visual motion consistencies. For feature-based detection, image alignment or frame registration is quite essential. The transformation of a different domain of images into a single domain is considered. Hence illumination, orientation and the size of the image play a vital role in this approach. This approach can be categorized as shape-based and colour-based methods. The procedure of shape-based or edge-based approach is filtering, enhancement and identification. The procedure works based on the Expected Maximization (EM) calculation of parameters in a random manner. In picture processing, an image acquisition device is used to capture the image, features are identified by the algorithm, putative factors are collected and objects are detected using HAAR classifier and Viola Jones framework (Neelima, Srikrishna, & Rao, 2017).

For template-based detection, the result is accurate provided that the template of the object with a high degree of precision is available. It can be further classified into fixed template matching and deformable template-based detection. In the fixed template, image is subtracted and matched by correlation. Deformable template matching is defined by the bitmaps describing the characters of the edges or shape of the object. It is applicable when the object varies between the rigid and non-rigid deformations (Kumaran & Reddy, 2017).

2.2 Object Detection and Recognition Techniques

Deep learning and machine learning are some of the approaches used in object recognition. These techniques learn to identify target objects in images using different approaches. Machine learning algorithms do not reply on models, they instead "learn" information from data directly. The performance of these algorithms improves with an increase in available samples for the learning process. This approach is suitable for problems where lots of variables and data is available but there is no existing model or equation. In general, machine learning algorithms fall into three categories namely-supervised learning, unsupervised learning and semi-supervised learning (Rosebrock, 2017). Deep learning is a flexible technology which has achieved a very high level of accuracy making it essential in applications such as driverless cars and speech recognition. This approach requires substantial computing power and a large amount of labeled data. Deep learning models are trained using neural network architectures and a large number of labeled data that learn features from data directly without manual feature extraction. Deep learning models have been proposed for several machine learning assignments such as Convolutional Neural Network (CNN), Deep Believe Network (DBN), and Deep Sparse Coding (DeepSC). The CNN technique is best suitable for image-based tasks such as face recognition and character recognition (Y. Ren, Chen, Li, & Kuo, 2018).