EE297 Intelligent Systems Project
Interim Report

Maynooth University
National University of Ireland Maynooth

Team 1

Frank Gallagher

Adam Duke

Artur Karolewski

James Oliver

TABLE OF CONTENTS

ABSTRACT

A signal is a function of one or more variables which conveys information on the nature of a physical phenomenon or system. However, signals can be corrupted by an unwanted signal (noise). Signal processing is the means by which noise is removed and information is extracted from signals. Processing signals provides us with information about the world around us. The physical nature of signals varies widely. The purpose of sensors is to capture signals of a given physical nature and translate them into electrical signals that computers can understand and analyse. Humans, and oftentimes machines, rely most heavily on audio-visual information for the perception of and interaction with their environment. For this reason, audio and vision processing are two primary signal processing fields. As the desire for intelligent devices grows so does the importance for Computer Vision, since vision is such an information laden signal.

Given a Logitech C270 webcam and a Raspberry Pi, we are assigned the task of designing a device which implements audio and/or vision signal processing. We decided to develop a vision system that analyses how many people pass a point of interest, how many of those people look towards the point of interest and how long they look for. The system could have several potential use cases but is designed with the intention of deploying it to analyse the attention given to an advertisement. In this report, we introduce our idea and the ethical considerations associated with it. A brief commentary on Person Detection and Pose Estimation, as theoretical bases of our proposed idea, is provided. The workings of the system are discussed in more detail along with potential implementation difficulties. Finally, we propose our plans for the next stages of development and provide some concluding remarks.

1 INTRODUCTION

1.1 PROJECT BRIEF

Design and implement an application based on audio and/or vision processing using a Logitech C270 webcam and a Raspberry Pi Model 3.

1.2 CHOSEN APPLICATION

A vision system that checks if a person looks in a particular direction. More specifically a system that analyses how many people pass a point of interest, how many of those people look towards the point of interest and how long they look for.

This requires the implementation of person detection and head pose estimation. Person detection is the process of determining the presence of a person in an image. This is necessary to count the number of people that pass the point of interest. The person needs to be detected before their head can be located for head pose estimation. Head pose estimation is concerned with describing the orientation of a person's head in a 2D image using a 3D coordinate system. This is required to determine whether or not the person looks at the point of interest. Extensive researched has been carried out in the fields of Person Detection and Head Pose Estimation. For example Paul Viola and Michael Jones propose the *Haar Cascade* approach for object detection in their 2001 paper "Rapid Object Detection using a Boosted Cascade of Simple Features" [1]. Naveet Dalal and Bill Triggs write about the *Histograms of Orientated Gradients* approach to Human Detection in their 2005 paper "Histograms of orientated gradients for human detection" [2]. The development of the field of Head Pose estimation is discussed in "Head Pose Estimation in Computer Vision a survey" [3].

Detecting people in images is a challenging task owing to their variable appearance and the wide range of poses they can adopt [2]. The ability to cleanly discriminate the human form in cluttered backgrounds and under difficult illumination yields a more accurate system. The variable appearance of people also poses challenges for head pose estimation as the system must be robust enough to reliably detect heads under various conditions, such as the presence of facial accessories like hats and glasses.

Our aim is to combine these two fields in one system, designed to be deployed in a customer analytics capacity. Using a camera and a Raspberry Pi running a smart algorithm to achieve this. Developing the smart algorithm using pretrained models from the OpenVINO toolkit which are optimized to run efficiently on the Raspberry Pi with a Movidius Neural Compute Stick. By developing the algorithm to run efficiently, we combat the issues inherent in Person Detection and Head Pose Estimation. Addressing these performance issues is especially important for our application, since our it requires the true measurement of people and their gaze.

1.3 ETHICAL CONSIDERATIONS

Our application requires video footage of people to perform its function. The intended use case of the system is customer analytics which means that it will be deployed in public places. Such footage is sensitive personal data and could potentially be taken without the intentional participation of the individuals filmed due to the public setting. Although Irish Law permits recording in public places, it would be considered ethical to notify people that recording is in progress, during both testing and deployment. The system is not designed to identify individuals and track their interest in advertisements to build a consumer profile. The system analyses the footage to detect human shapes and determine the oriented of the head. Once the footage has been analysed and the results logged, the footage is deleted. The only data stored is an anonymous tally of the number of people that pass the point of interest, how many look at it and how long for.

Since testing will be conducted on a University Campus, which is not considered fully public, permission will be sought to record for testing purposes and when testing is in progress people will be notified of the recording.

2 THEORETICAL BACKGROUND

In this section we discuss certain theoretical aspects relevant to our chosen application. We are attempting to develop a system that analyses the number of people that pass a point of interest, how many of these people look in the direction of the point of interest and how long they look for. This application has two main constituent components:

(i) Person detection.
(ii) Pose estimation.

2.1 PERSON DETECTION

Person detection is a form of object detection where the object class is 'Human'. Object detection is the process of determining the presence of predefined type of object in an image, typically without the intentional participation of the detected person. This task involves determining both the presence of the object and its location in the image. The image can be a typical colour image or an infrared image. Common applications of person detection are search and rescue, surveillance and customer analytics.

Approaches to person detection fall into two primary categories: (i) classical (ii) modern. In general, both approaches attempt to find common shapes in images, often utilising machine learning in an attempt to avoid hand crafting features and hyperparameters, instead relying on huge amounts of training data. However, the main distinction is that modern approaches rely more heavily on *Deep Convolutional Neural Networks* to achieve more efficient performance and more accurate results.

Classical approaches include *Haar Cascade* [1] and *HOG (Histograms of Oriented Gradients)* [2]. The *Haar Cascade* approach "combines increasingly more complex classifiers in a "cascade" which allows background regions of the image to be quickly discarded while spending more computation on promising object-like regions. The cascade can be viewed as an object specific focus-of-attention mechanism, which unlike previous approaches provides statistical guarantees that discarded regions are unlikely to contain the object of interest" [1] (*Figure 1* below shows the *Haar Cascade* approach being used for person detection). *HOG* is a type of feature descriptor. The purpose of a feature descriptor is to generalise an object so that it can be detected when viewed under different conditions [4]. Drawbacks associated with these approaches include missed detections, false detections and duplicate detections [5]. Such problems arise as a result of the variable appearance of people and the wide range of poses they adopt. These problems are more likely to arise in areas with significant human activity, such as high streets and shopping centres, which are probable target locations for our system. Issues with missed detections and false detections are especially problematic for a system with the sole purpose of gathering accurate detection numbers.

The use of *Deep Convolutional Neural Networks* attempts to address these problems by enhancing system performance. A brief description of neural networks is provided in section 2.3 below.

Figure 1 - Person detection using Haar Cascade approach [6]

2.2 POSE ESTIMATION

3D pose estimation is the problem of determining the 3D orientation of an object from a 2D image. Human pose estimation is the process of estimating the configuration of the body from an image. Head pose estimation is concerned with inferring the orientation of a human head from digital imagery [3]. Head pose can be described by specifying the three rotational angles (Pitch, Yaw and Roll) about the three corresponding translational axes (X, Y and Z axes), shown in *Figure 2* below. This involves first determining the location of the head in the image and establishing the origin of the 3D coordinate system, which requires facial detection. This is where the difficulties of head pose estimation lie, since a non-pose-specific face detection system is needed. Otherwise, a 'chicken-egg' problem arises: the location of the head is necessary to determine its pose, while the pose is necessary to locate the face [7]. Another issue with head pose estimation is that a head pose estimator must be able to display invariance to biological factors such as facial expression, skin colour and the presence of accessories such as hats and glasses [3]. There are a number of approaches to head pose estimation, including *Detector Array Methods* and *Geometric Methods*. The *Detector Array Methods* involve employing machine learning and neural networks to train a series of head detectors each configured to a specific pose. *Geometric Methods* use the position of facial features such as eyes, mouth and nose to determine head pose from their relative configuration [3].

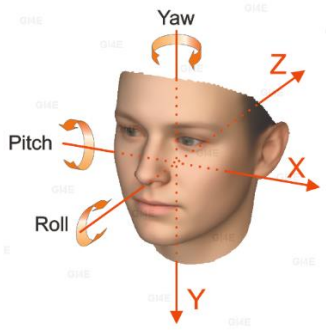

Figure 2 - Definition of translation and rotation axes [8]

3 IMPLEMENTATION

This section outlines the general structure of our system as well as the limitations of the hardware and the issues associated with these limitations.

3.1 SYSTEM STRUCTURE

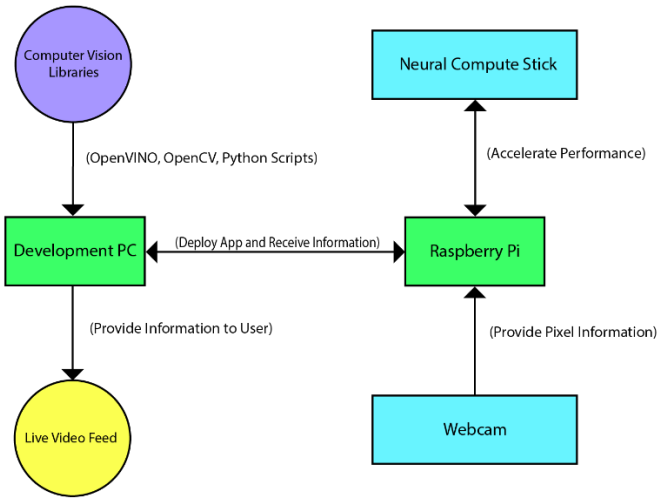

Figure 3 - Flow diagram of general system structure

The system (as depicted in *Figure 3* above) consists of a webcam, Raspberry Pi, Neural Compute Stick and a development PC. The software for the system is composed on the development PC using computer vision libraries, such as OpenVINO and OpenCV. This is then built into an application and deployed to the Raspberry Pi. The app runs on the Raspberry Pi and receives the video feed from the webcam as input. The Neural Compute Stick accelerates performance. The Raspberry Pi is remotely connected to the development PC. This allows the user to view the live video feed from the webcam and the results of the analysis while the system is running, as it removes the need for the Raspberry Pi to be connected to a monitor.

More specifically, the webcam is sending pixel data into the app running on the Raspberry Pi. This pixel data is passed into a person detector to determine whether or not a person is present. If a person is present, a people counter is incremented and the head pose estimator is activated. The system then checks if the detected person was looking towards the point of interest by comparing the head pose angles against predefined angle ranges. The time duration that the head pose angles lie within these ranges is considered the length of time the person looked at the point of interest.

The system structure outlined above implements the Intel Distribution of the OpenVINO toolkit. This is designed to optimise the performance of Computer Vision applications on Intel hardware, such as a Movidius Neural Compute Stick. The OpenVINO toolkit includes pretrained models for person detection and head pose estimation and therefore negates the need to acquire sets of training data and expend vast amounts of time and effort training a neural network.

3.2 Hardware Limitations and Implementation Issues

The Raspberry Pi has 1GB RAM and a clock speed of 1.2GHz. The Logitech C270 webcam has a resolution of 3MP and operates at 720p with a fixed focus beyond 40cm. It doesn't feature a wide angle lens, nor does it respond well to poor light conditions. It can achieve an output of between 10 and 20 fps.

The system must be capable of detecting people walking at a normal pace ($1.5ms^{-1}$) and detecting the movement of a person's head. The an average adult male can move their head forward and backward (pitch) from $-60.4°$ to $69.6°$, bend their head right to left (roll) from $-40.9°$ to $36.3°$, and rotate their head (yaw) from $-79.8°$ to $75.3°$ [3]. These abilities are limited by the speed of the Raspberry Pi, the resolution of the webcam and the fps of the video feed. The absence of a wide angle lens means that a person will be in frame for a shorter time, thus decreasing the time the system has to detect the person and estimate their head pose. The fps and resolution of the webcam may also limit the systems ability to detect glances (minimal head rotation in terms of degrees and time). The purpose of the Neural Compute Stick is to enhance system performance and reduce the effects of the hardware limitations.

4 CONCLUSION

In this report, we propose an idea for a vision processing based application. Ethical considerations for such the application are provided. Following this the constituent theoretical aspects of the app, Person Detection and Head Pose Estimation, are briefly discussed. The proposed workings of the system are then outlined along with the hardware limitations and associated implementation issues.

The proposed system is a smart camera that analyses how many people walk past a point of interest, how many look towards the point of interest and how long they look for. This involves detecting the presence of a person in the camera frame, locating their head and estimating the direction of their gaze. The average walking pace of a person and the range of head movement must be considered.

The system structure outlined in the previous section implements the OpenVINO toolkit. This approach involves using pretrained models, included in the toolkit, for person detection and head pose estimation. In the coming weeks the team is going to explore the use of the OpenVINO toolkit to develop our proposed application. First sample OpenVINO apps will be built. We will then develop our own algorithm, tailored to our specific application, using the sample apps as a guide.

REFERENCES

[1] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," 2001.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, 2005, vol. I, no. 3, pp. 886–893.

[3] E. Murphy-Chutorian and M. Manubhai Trivedi, "IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE 1 Head Pose Estimation in Computer Vision: A Survey," 2008.

[4] C. McCormick, "HOG Person Detector Tutorial · Chris McCormick," *mccormickml.com*, 2013. [Online]. Available: http://mccormickml.com/2013/05/09/hog-person-detector-tutorial/. [Accessed: 13-Mar-2019].

[5] M. Vidanapathirana, "Real-time Human Detection in Computer Vision — Part 1," *medium.com*, 2018. [Online]. Available: https://medium.com/@madhawavidanapathirana/https-medium-com-madhawavidanapathirana-real-time-human-detection-in-computer-vision-part-1-2acb851f4e55. [Accessed: 13-Mar-2019].

[6] M. Vidanapathirana, "Haar cascade based Human Detection," *medium.com*, 2018. .

[7] J. Sherrah, J. Ong, and S. Gong, "Estimation of Human Head Pose using Similarity Measures," *Queen Mary University of London, School of Electronic Engineering and Computer Science*, 2000. [Online]. Available: http://www.eecs.qmul.ac.uk/~sgg/pose/. [Accessed: 13-Mar-2019].

[8] M. Ariz, J. J. Bengoechea, A. Villanueva, and R. Cabeza, "A novel 2D/3D database with automatic face annotation for head tracking and pose estimation," *Comput. Vis. Image Underst.*, vol. 148, pp. 201–210, Jul. 2016.