Computing Science and Mathematics
University of Stirling

Contextual Understanding and Object Intent Detection

Daniel Rooney
2535156

Supervisors: Dr. Deepayan Bhowmik,
Christopher Dickson(Thales UK)

Dissertation Outline

27th October 2020

Contents

1 Introduction **4**
- 1.1 Background and Context . . . 4
- 1.2 Scope and Objectives . . . 5
  - 1.2.1 Scope . . . 5
  - 1.2.2 Objectives . . . 5

2 State-of-The-Art **6**
- 2.1 Previous Work . . . 6
- 2.2 Deep Learning . . . 6
  - 2.2.1 Neural Networks and the Multi-Layer-Perceptron . . . 7
- 2.3 Activity Recognition & Object identification . . . 11
  - 2.3.1 Convolutional Neural Networks . . . 11
  - 2.3.2 Supervised & Unsupervised learning . . . 13
- 2.4 Semantic Scene Segmentation . . . 13
  - 2.4.1 Part & Key point Recognition . . . 14
- 2.5 Contextual Awareness . . . 15
- 2.6 Technologies & Hardware . . . 15

3 Problem description and analysis **16**
- 3.1 Problem Description . . . 16
- 3.2 Problem Analysis . . . 16
  - 3.2.1 Contextual recognition and awareness . . . 17
  - 3.2.2 Intent Detection and subsequent identification . . . 17
  - 3.2.3 Situation evaluation and notification . . . 17
- 3.3 Constraints . . . 17
- 3.4 Requirements . . . 18
  - 3.4.1 Functional Requirements . . . 18
  - 3.4.2 Additional Requirements . . . 18
- 3.5 User Stories . . . 19
- 3.6 Assumptions . . . 20
- 3.7 Approach . . . 20
  - 3.7.1 Agile Development . . . 20

4 Project Plan **21**
- 4.1 Overview . . . 21
- 4.2 Timeline . . . 21

4.3 Deliverables 22
4.4 Intended Result 22

5 References **23**
5.1 Figure References 24

List of Figures

2.1 An artificial neuron, where x is an input.[A] 7
2.2 Basic diagram of a three layer Multi Layer Perceptron.[B] 7
2.3 Sigmoid equation and its graphical form.[C] 8
2.4 Sigmoid & ReLu Functions respectively.[D] 10
2.5 Cost function equation, where y(x) is the approximation of all weights biases where x is the input, b are biases, n the number of raining inputs and a which denotes the output vectors of the inputs.[E] 10
2.6 A representation of gradient descent with routes to local minima.[F] 11
2.7 Diagram showing the architecture of a CNN.[G] 12
2.8 Max pooling using a 2x2 filter and stride length of 1.[H] 13
2.9 An example of scene segmentation separating two elements(classes) of a scene.[I] 14
2.10 Scene segmentation applied to a busy street.[J] 14

4.1 Gannt Chart Roadmap 21

Chapter 1

Introduction

The field of computer vision at present is one of the biggest areas of artificial intelligence right now. The use of neural networks and machine learning algorithms have allowed computers to recognise, detect, track and classify objects with relative accuracy in real time, as well as break up images and scenes into seperate elements for better understanding and evaluation. However one thing that the field lacks is the ability to understand the context of a situation, and potentially preempt what may occur based on what the computer can see. This is what I aim to achieve in this project. To accomplish this I propose a split system, using object detection and tracking to classify and monitor an object, and scene segmentation to break up the environment surrounding these objects - that can be used to evaluate and produce a context. Using both environmental variables and information about an object and it's behaviour, and then by evaluating them against a set of pre-defined rules, will allow the system to "preempt" and notify the user what an object will do next, and what it's potential intent may be. The system could potentially be applied to areas of safety and accident prevention, threat detection and situation monitoring over a wide range of contexts.

1.1 Background and Context

Everyone has seen videos in the news or online of situations that, if someone intervened only minutes before, could have had a different outcome. Unfortunately some of these situations are easily preventable if someone could have recognised a problem or incident. However we're only human and sometimes we miss things. But what if there was a computer vision application that could detect suspicious behaviour or actions and let someone know ahead of time, potentially deescalating a situation or potentially save a life. Situations like attempted suicides, dangerous traffic or even terrorism incidents, could potentially be identified before they occur, based on a set of contextual factors. For clarification, a good example of this would be someone lingering by the side of a bridge for a long period of time. There are certain questions we could ask, such as; are they alone? Are they pacing? Are they wearing work clothes? Do they have any possessions around them? Have they been in the same place for a long time and are they sitting on the edge? The answers to these questions could determine whether we check on that person, or leave them alone. That in essence, is what this project hopes to achieve, to let the user know whether they should be aware of something going on and what the possible intentions are of that person/object, before they are carried out. This is a multi-purpose solution to a set number of problems, that could be solved or

improved by a piece of intuitive software.

1.2 Scope and Objectives

1.2.1 Scope

The scope of the project can be split into 3 main areas:

- Object Detection and Tracking - The software will utilise current libraries and frameworks to detect and classify an object. It will also track an object and calculate it's speed and location within a scene. It will potentially be able to group objects together if they are within close enough proximity to each other.
- Semantic Scene Segmentation for contextual understanding - The software will use computer vision methods such as a convolutional neural network model to classify each pixel/ or group of pixels within a scene, to split it up into several elements that can be evaluated together to determine a visual context.
- Evaluation Algorithm - An algorithm that will receive inputs from both processes, evaluate the classified object(s) and their movements, and the environmental elements against a pre-determined rule-set dependent on use case. It will then determine an objects intent; and relay this back to the user.

1.2.2 Objectives

The objective of this project is to utilise both object detection and tracking methods alongside semantic scene segmentation (to contextualise a scene using present elements and location) to produce a better machine understanding of situational context in real time and notify a user. To be used in accident prevention, asset protection, health and safety protocols or anywhere it could be applied in a relative industry, to optimise current hard-written protocols and hopefully reduce overhead cost where applicable.

Chapter 2

State-of-The-Art

2.1 Previous Work

Previous examples of work based around the concepts used within this concept, all use elements of the following methods explained and analysed in this section. Current state of the Art research has been carried out at Microsoft Research, producing a human pose-estimation concept [15], this concept is able to;with relative accuracy estimate the following pose by a human in a sequence of moves.However it does not seem to estimate entire groupings of movement, but rather just the separate poses made between two periods of time. The work of Thomas.M.Strat of SRI International aimed to identify, with moderate success the contextual information received from a visual input using the CONDOR Algorithm,[14] however context differs from situation to situation, so some sort of external analysis could be implemented to recognise certain situations. Due to the extensive amount of work done using deep learning methods for object and activity recognition, I cannot narrow down any specific literature to analyse, as it is safe to assume from what my research has lead me to read, listen and watch, advanced object identification is possible, and the concepts used could easily be applied to this project as they are well documented and widely used. The fundamental concepts used in computer vision are discussed in the section below.

2.2 Deep Learning

Deep learning is the process and implementation of functions that can imitate the human brain and it's thought processes. These include, but are not limited to; decision making, perception, spacial awareness and recognition. Deep learning is a crucial area of machine learning; in modelling the human brain with what are called "Neural networks", we can start to process data and create patterns to improve machine decision making, which is the biggest component of deep learning. Furthermore, deep learning enables networks to learn things by themselves, unsupervised from lots of data, something that would take humans much much longer, and this data may not be structured or laid out in any sense, making it better to use intelligent neural networks, giving them an essence of "Independence".
In recent years, we have seen an increase in deep learning applications and huge advances in what we can achieve with them, it truly is becoming a pillar of the "digital age".

2.2.1 Neural Networks and the Multi-Layer-Perceptron

As stated previously, Neural Networks are algorithms that are inspired by the human brain. These networks, like the human brain, make use of neurons(also called nodes); artificial neurons to be exact. These neurons work and process data by taking several binary inputs, and then producing an output. [Figure 2.1]

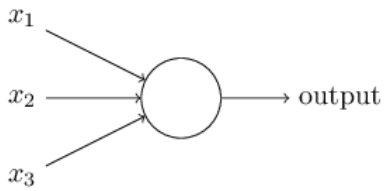

Figure 2.1: An artificial neuron, where x is an input.[A]

The multi-layer perceptron, is a class of feed forward neural network, it is also commonly known as a "vanilla network".[10] It is organised into three segments, an input layer, a hidden layer(s) and an output layer, feeding data through it until the network produces a result. [see fig 2.2]

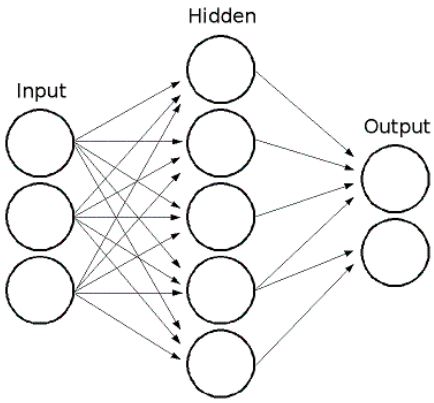

Figure 2.2: Basic diagram of a three layer Multi Layer Perceptron.[B]

Each neuron is given an activation value, in MLP, this is a value between 0 and 1 (for simplicity in this explanation, we assume the network is breaking down a greyscale image), with 0 usually being assigned to completely black, and 1 to white, with a range in between. When these neurons take in an input, their value will correspond to the colour intensity of whatever section or pixel the neuron is observing. Furthermore the activation's in the input layer, determine the activation's in the next layer.[4] The connections between these layers are called weights, weights are numbers which can be used to tweak and change how the neural network behaves, following with the example of a greyscale image, we can assign a positive weight to

neuron connections with an activation value closer to 1 (or closer to white), and a negative weight to an activation value closer to 0 (or closer to black), with the weights differing depending where they correspond to each neurons activation value on the scale between 0 and 1. Therefore the final activation function of each neuron is how positive the sum of the relative weighted value is. This calculation is carried out on every neuron in the network, which gives us the value in where the neurons will activate. This is called the activation function, to break it down, the product of each neurons weight and activation is taken, and then added together in series[10]:

$$(a_1w_1 + a_2w_2 + ...a_nw_n)$$

However when calculating a weighted sum, we want to be able to condense it down into our range of 0 to 1, this is achieved using the Sigmoid function. In a Sigmoid function, negative inputs are closer to 0, and more positive inputs are closer to 1, with a steady positive increase around the input of 0:

$$\sigma = \frac{1}{1 + e^{-}(x)}$$

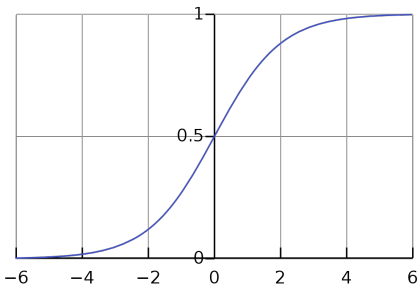

Figure 2.3: Sigmoid equation and its graphical form.[C]

We can use this function to to determine that the activation value is just a measure of how positive the relevant weighted sum is, which when coupled with the relevant series of weight and activation products, produces:

$$\sigma(a_1w_1 + a_2w_2 + ...a_nw_n)$$

However for the purpose of tuning a neural network, one may not want a neuron to activate if its value is only bigger than 0, to change this, we use a Bias, which can be any given

value, to amend the function and cause the activation to occur when the result is larger(or smaller) than the new Bias inherent value:

$$\sigma((a_1w_1 + a_2w_2 + ...a_nw_n + B)$$

This expression can also be written as a set of vectors and a matrix,however in this example it shall be condensed down, for ease of use when applying it to real world applications:

$$A^1 = \sigma(Wa^0 + B)$$

Therefore, in conclusion, the weights of each connection between neurons indicate which pixel pattern a given neuron is recognising, and the Bias gives the principle value that the neuron is required to reach before it activates. Each connection between neurons, has its own weight associated with it, and each neuron has its own Bias.[4][10] Both of these variables are *Hyperparameters*, large sets of parameters that if organised a certain way yield the best results, there is no specific magic formula to achieve this ideal state,[5] it is a case of taking an educated guess and using gradient descent to determine where to go next (this will be explained later). When training a neural network, its "learning" is essentially the process of finding the best weights and Bias that correspond with a higher accuracy rate.

Rectified Linear Unit Function (ReLu)

It is important to note, that in relation to modern, state of the art CNN's and MLP, that the majority of modern neural networks don't make use of the Sigmoid Function anymore, as it is considered outdated. Most state of the art networks use the ReLu function. This is due to the "vanishing gradient function",[2] where hyperbolic tangent functions (like Sigmoid), tend to vanish as network complexity increases, due to the ever increasing amount of layers in state of the art artificial neural networks. ReLu allows us to train these networks much faster, as it is a much simpler function to implement, and in doing so, the accuracy of these modern networks has improved too. Below is a comparison between the ReLu function and the Sigmoid function[fig 2.4]:

The Cost Function & Gradient Descent

The cost function is the average of every training examples cost. The cost of a training example is the squared differences of the input values and output values (the product of activation values, weights ad biases). The difference being what the neural network has recognised the input as (in the form of a numerical value) and what the actual input value

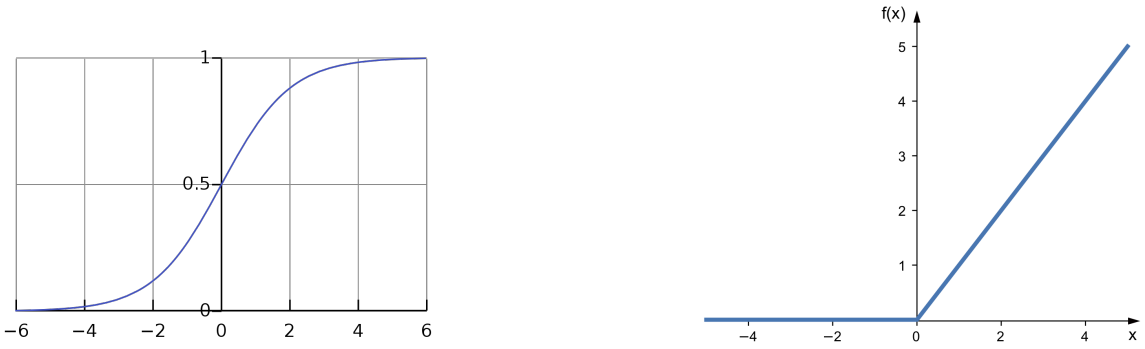

Figure 2.4: Sigmoid & ReLu Functions respectively.[D]

is. This product can be taken as an indicator of the performance of the neural network. Therefore, the lower the "cost", the lesser the difference in values, and the more accurate the output and vice versa. In training, the cost is function is the average of all of these individual cost values, to minimise the value of the cost function, we can use gradient descent.[10][4]

$$C(w,b) \equiv \frac{1}{2n} \sum_x \|y(x) - a\|^2.$$

Figure 2.5: Cost function equation, where y(x) is the approximation of all weights biases where x is the input, b are biases, n the number of raining inputs and a which denotes the output vectors of the inputs.[E]

Gradient descent is the process of calculating the gradient of the cost function, which in turn gives us the direction of steepest ascent within a graph, if we simply make this value negative, we then get the direction of steepest descent:

$$-\Delta = C(w,b)$$

We can use this to determine which direction to go in next, again there is no rule of which input we must start with before we apply this method, it is just a case of constantly refreshing (moving to a different position) the position of a cost function until we reach a *local minima*. A *local minima* being the smallest possible value for a cost function we can achieve within a given area. By finding these *local minima*, we achieve a lower cost function, hence improving the accuracy of the neural network. Within every set of data, we also have a *global minima*, which is the lowest possible value of the cost function, which is ideal, however it proves exceptionally difficult to find, as with increasing inputs, the more calculations we must do with

different positions to achieve it,[4] and this takes a long time. It is important to note that the greater the "refresh rate" the greater chance of overstepping a *local minima*, and vice versa for a lower rate, where gradient descent progresses too slowly. In [figure 2.6] below a representation of a function with multiple inputs with the cost function represented above in a separate plane, can help visualise the process of gradient descent.

Backpropagation

Backpropagation is the process of manipulating weights within a neural network ever so slightly to achieve the lowest value of the cost function. This process takes extensive calculation and will be built upon at a later stage in the project.[1]

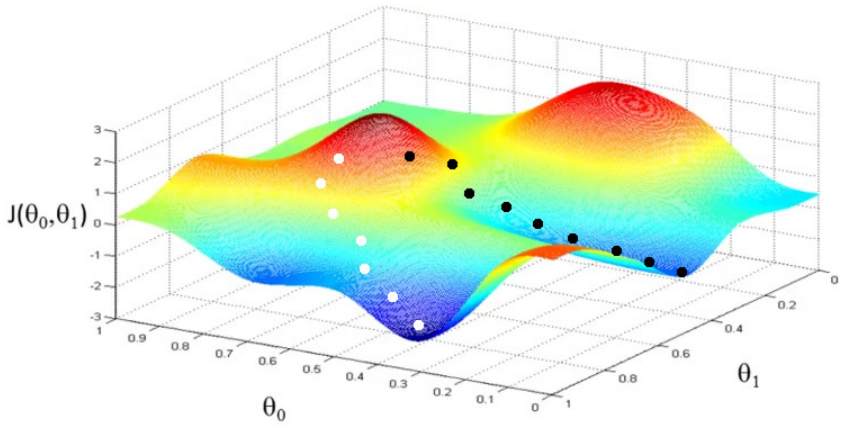

Figure 2.6: A representation of gradient descent with routes to local minima.[F]

2.3 Activity Recognition & Object identification

2.3.1 Convolutional Neural Networks

A Convolutional Neural Network , is a class of neural network that is made up of convolutional layers (most modern CNN's also have non-convolutional layers, like the MLP), convolutional layers, also known as the *kernel*, perform a transformation known as the convolution operation. To understand this, we must first understand that each convolutional layer is made up of filters, these filters are capable of detecting patterns (like the MLP, but more thorough). This makes them well suited to image/object recognition, as a set of filters, each one detecting a different form of pattern, encased in a single convolutional layer, enables a CNN to greater break down the elements of a visual input.[12] For example, one filter may recognise edges, or texture. The deeper the CNN, the more complex and specific these filters become, they may start to look for unique patterns instead, such as the eyes of a human, a type of camouflage pattern or a certain animal.

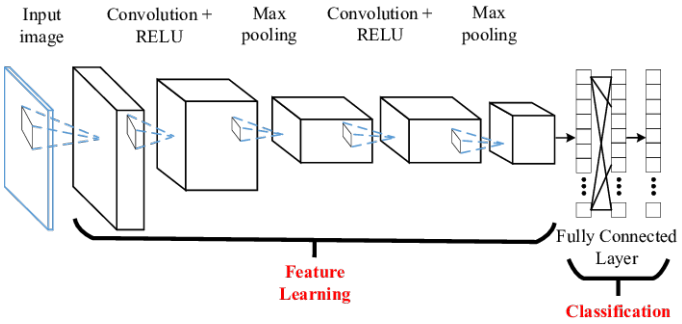

Figure 2.7: Diagram showing the architecture of a CNN.[G]

The essence of this process is essentially breaking down every element, classifying them in different ways, and then adding all of this data together to produce an output for the next layer, or the final output. [3] The way in which a CNN accomplishes this is known as convolving; each filter is made up of relatively small values within a matrix, and these values are then initialised with random numbers (much like in an MLP), these matrices (or "filters") stride over an image in a certain direction (left, right etc),[3] taking the dot product (sum of all multiplied matrix values) of the input value and the random values within the matrix, and producing an output value.

Pooling Layers

Pooling layers are used in CNN's to reduce the size of a convolved feature[12]. This makes the system more efficient and less resource heavy, which will be an important element to take into account when using an embedded system.

The two most common types of pooling are *Max Pooling* and *Average Pooling*. However average pooling is used substantially less than max pooling, therefore the focus of my research is on max pooling.
Max pooling is the process of breaking down an input/feature into regions and taking the maximum activation value of that region (the higher the activation, the greater chance a pattern has been recognised)[11,12], and producing a condensed output from the original input. This improve help performance in CNN's, as it disregards any data that has not hit its activation value, however, due to it's disregard of certain data (also known as down sampling), this could potentially decrease accuracy and the models score. An example of max pooling can be seen below[figure 2.8]:

Max pooling makes use of two hyperparameters, the *filter size* and the *stride length*, these are used to determine the highest activation value in each region, as can be seen above in [figure 2.8].

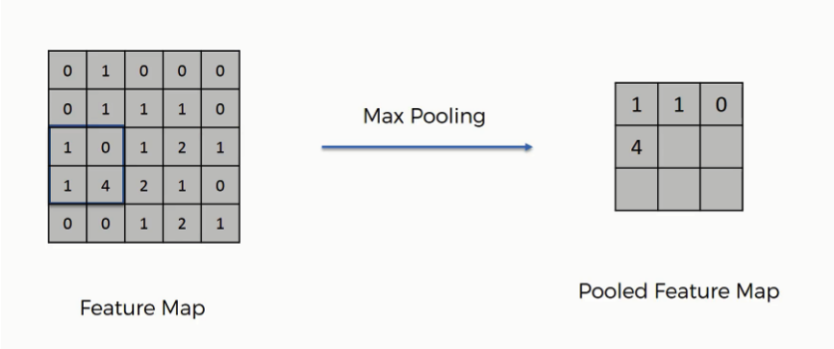

Figure 2.8: Max pooling using a 2x2 filter and stride length of 1.[H]

Fully Connected Layer

A Fully Connected layer, performs classification, until now, both the kernel layers and the pooling layers have carried out *feature learning*. In a fully connected layer, an input is first flattened, this is the process of converting an image/frame into a column vector[11], we then apply backpropagation[1] to this input before an output is produced. This is done for every sequence of training to produce a more accurate output.

2.3.2 Supervised & Unsupervised learning

Supervised Learning is the process of labelling the datasets that we use in training neural networks, it is done when we are performing classifications. Furthermore, in supervised learning, the principle of *ground truth* is applied, this means we know what output we would like the neural network to produce.[13]
Unsupervised learning is the process of not labelling datasets, in which the neural network simply outputs data and classifies it as what it "thinks" that data is.[13]

2.4 Semantic Scene Segmentation

Semantic scene segmentation is the process of taking a visual input ((image, video etc.),and labelling each pixel(or group of pixels) within the input and classifying them. It is essentially splitting up an image/video into components [see figure 2.9]:

Semantic scene segmentation can be achieved by labelling and categorising elements within images/videos and using them to train a CNN, this CNN can then take a visual input and segment the "scene" into it's respective classes based on it's previous training with datasets.

A good example of modern Semantic Scene Segmentation is state of the art autonomous automobiles. These vehicles can use CNN's and SSS to label and classify data within their field of view, and use that information to detect and recognise things like lanes, people, crosswalks

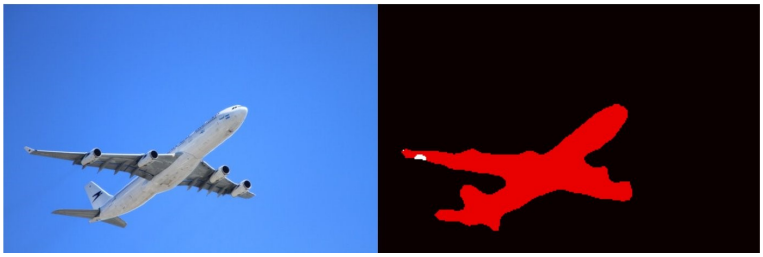

Figure 2.9: An example of scene segmentation separating two elements(classes) of a scene.[I]

and other vehicles. These state of the art applications of SSS are already highly accurate, but lack the contextual awareness to react to a situation that they may not have encountered.

Figure 2.10: Scene segmentation applied to a busy street.[J]

2.4.1 Part & Key point Recognition

Part and key point recognition is the process of breaking down an image/frame and identifying certain areas with points. These points can then be connected to produce a point net, which can form a partition of an element in an image, with every point mapping to its previous point when moved to a different location. This can be used in conjunction with pose & action analysis to track movements and make predictions about where an object will move next within a given frame.[8]

2.5 Contextual Awareness

The contextual awareness of a system that employs computer vision techniques could be defined as follows: *"The ability of a system to recognise and classify contextual elements within prevalent within it's input data to make a prediction on the situation that it finds itself in."*. Therefore it is essential to break down these contextual elements and categorise them, and in doing so the system will be able to pull and compile the classified elements to produce an output. This has already been coined as context-based vision (CBV). Thomas M. Strat [14], has conducted extensive research on this topic, and defines three types of specific context, they are as follows:

- **The Physical Context** - The physical context of a situation (or image, video etc.) is actually a well known category of context, not specific to any previous personal work. Physical context is the generic information we can define from a scene, as previously used, a woman walking across a street is not very intuitive to the context of a situation. However, if we define the physical elements of this scene, a better understanding can be gained, such as the surroundings, like buildings, foliage, moving objects and dynamic elements of the interpreted image.[14]

2.6 Technologies & Hardware

Software

Several technologies and libraries exist that aid in the creation and testing of deep learning and computer vision applications, such as openCV, openCV Keras, TensorFlow, pyTorch and torchbearer. For this project the frameworks and libraries to be used are openCV with Keras and Tensorflow, as these are more suited to the production standard of this solution, and easier calibration and testing on the finished product. The main programming language to be used is python, due to it's ease of use, modular structure and large community for support.

Chapter 3

Problem description and analysis

3.1 Problem Description

Many existing platforms (including the one in current use at Thales UK) are capable of identifying objects and labelling them when they are present in the camera view. However, these objects/individuals are all treated as a person of interest and their intent is not known. For example, an old person walking across a street and minding their own business is not considered a person of interest in this scenario. However, an individual peering from the top window of a building with an object in their hand(s) may incur a point of interest to the user. The classification of a potential threat can only be positively determined by context, or a sequence of behaviours that may amount to some sort of action(s) that could be deemed a threat based on previous datasets. Therefore the problem we are faced with is the inability current software to be able to understand the context of a situation and the several factors involved that may lead to a reasonable conclusion of what is happening or about to happen at that point in time.

It is important to note that when "existing platform" is mentioned, it is not a code base that I will implementing my work on top of, it is the assumption that the company platform used currently by Thales already identifies certain objects. Therefore I will be implementing my own version of this platform (which is already very commonplace in computer vision applications), and building upon it so that it may carry out the intended requirements that have been outlined for this project.

3.2 Problem Analysis

Breaking down the problem is somewhat difficult, as the individual sections of the proposed solution are inevitably interlinked. However, working on the assumption that the base of the solution is already available and already has the ability to recognise objects and/or faces, label them and keep a recollection, the problem can be broken down into three explicit task areas:

3.2.1 Contextual recognition and awareness

The current systems have no method of identifying and categorising the context around an object or individual of interest. Therefore, the system may only make a single assumption with no real basis, as right now it only acts as an identifier. For example, an individual on a rooftop can safely be categorised as a "point of interest" whereas, a child playing with a football in a park is most likely not a point of interest. However currently the system categorises them both as a "point of interest", even if they are not, as it has no contextual environment data to come to a conclusion in it's identification. This could lead to potential false alarms, incorrect analysis and a convoluted line of sight if all objects are identified under the one identifier.

3.2.2 Intent Detection and subsequent identification

The current systems at present have no mechanisms of detecting object intention, and its subsequent identification/categorisation. An example of intent detection would be as follows: *"An individual moving at a fast pace within a cameras field of view, suddenly stops, behind a wall (Wall being a potential contextual identifier). This individual then reappears from the same location and raises their arms up to their line of sight, with a long, black object in hand."* With a predetermined understanding of human pose sequences, the new system would identify the object as a firearm or blunt object, and the sequence of movements by the individual identified as a potential firing stance or attack. The system then labels this within the cameras field of view, logs it, and raises it as a point of concern and/or a threat. This is the most challenging problem area, as it will require an in depth analysis of object identification, pose and activity recognition, as well as several thousand visual data-sets to train a neural network(s) to identify all of these components, log them and then evaluate them together to produce an accurate conclusion and notification. Furthermore, accuracy and error potential is still always an issue.

3.2.3 Situation evaluation and notification

The current methods of notification are simple box labelling methods.This does not provide any intuitive information about identified objects or the situation. Ideally the solution will have an event log, corresponding to the auto labelling boxes with all the concluded contextual and intent information. Furthermore the conclusive identifiers (threat, no threat and potential threat), will be displayed with the on screen labels within the line of sight. This part of the problem is better defined as a non-functional requirement, however it is just as important, as it is the point of contact with the operator. (GUI, notifications, labelling etc.).

3.3 Constraints

Below are a list of project constraints that the solution is limited by:

- **Embedded System (Technological Hardware Cost)-** The software solution (system) is intended to run on an embedded platform, with real-time computing constraints.Therefore it is essential that the system is able to run and perform it's tasks as part of an embedded system, and therefore cannot be too "heavy", as that would require a significant increase in computing power or a change to surroundings systems. The

proposed platforms for this project are the Nvidia Jetson TX2 and the Nvidia Jetson TK1.

- **Timeline-**An inevitable constraint, this project has a time period of around eight months from start to finish. Therefore a full production implementation is likely not possible. With how extensive the project is, several factors may have to be reduced, such as the amount of categorisation classes for events, the amount of data-sets used to train the Neural Network and several non-functional requirements may not be implemented entirely.

- **Development Environment-** The technological constraints of the machine(s) that the project is to be developed on must be taken into account. Especially towards completion of the project, other platforms with greater computing power may have to be used to run the demo during testing and debugging.

- **Skill-** The technical skills of myself and knowledge of the theory behind the project may be limited , and may result in time being lost due to extra reading, practice etc. This is of low concern.

3.4 Requirements

3.4.1 Functional Requirements

- **"Solution must run on an embedded platform"-** The solution must run on any specified embedded platform for use with other systems, therefore it must not be technologically heavy, and optimised for this requirement.

- **"Must be able to identify and categorise the context of a situation around an object"-**The solution must be able to identify the context of a situation around an object using environmental elements, for example: *Secluded area + Small Window + High building + Main road = potential threat.* Based on previous training using visual data-sets, the solution must come to a conclusion by using any identified environmental elements.

- **"Must be able to detect an object and classify it's intent"-** Using pose and action analysis, activity recognition and object identification, the solution must come to a conclusion of the objects intentions based on these variables.

- **"Must notify the operator of an objects presence and classify it"-** As above, the solution must identify an object. Furthermore it must produce an on screen notification that can be seen and read by the operator(eg. threat, no threat box label).

- **"Must be able to identify several objects at once in a given frame"-**Must be able to track and identify multiple objects at one time within any given frame.

3.4.2 Additional Requirements

- **"Must be able to compile an object log, with process information and conclusion"-**Compilation of an object log, including all processes and assumptions as well as additional process information that relates to the formed conclusion.

- **"Must provide an accuracy rating to produce notification"-** Must provide an accuracy rating when identifying an object and its intent, e.g. 0%-60% = Low accuracy, 60%-80% = Medium accuracy and 80%-100% (99.999) = High accuracy .
- **"Must be able to identify and remember previously identified objects"-**Must be able to identify faces and objects of interest, for example pre-loaded images of objects of interest, as well as previously identified objects that have not been pre-loaded.
- **"Must be convenient and easy to use"-** Must be easy to use, read and intuitive, within a number of environments.

3.5 User Stories

I have compiled some potential user stories that have been constructed from the given requirements. These user stories will be graded with a scope and importance:

User Story	Scope	Importance (scale of 1-5, with 1 being highest and 5 lowest)
1	Large	1
2	Large	1
3	Large	1
4	Medium	3
5	Medium	1
6	Medium	3
7	Large	2
8	Small	4
9	Small	1

User Story 1 - *"As an operator I want to be notified when an object of interest is in my field of view so I can act accordingly."*

User Story 2 - *"As an operator I want to be given an accurate intent classification and context evaluation, so I can better understand objects within my field of view and around me."*

User Story 3 - *"As an operator I want to be given prior warning on detected objects and track them within my field of view, so that I may monitor them accordingly."*

User Story 4 - *"As an operator I want to be able to upload visual data to the system (object,individual,face etc.) so that they can be immediately identified if seen."*

User Story 5 - *"As an operator I want to be presented with simple,concise accuracy ratings and classifications in the form of an easy to read on screen label/tag."*

User Story 6 - *"As an operator I want to be able to view a log of all events and identified objects within a given time period, with each event given it's own time tag."*

User Story 7 - *"As an operator I want to be able to identify multiple objects within the frame(s) of view so that I do not miss any myself."*

User Story 8 - *"As an operator I want to be able to view and add a different range of categories(classes), so that I can better understand a situation, whether it is present or future."*

User Story 9 - *"As an operator I want to be able to use the solution on an embedded platform so that I can implement it in tandem wit my own external systems."*

3.6 Assumptions

Below are a set of assumptions for this project that may contribute to it's overall success. These assumptions may change over time, but as of this time of writing before I begin the implementation of the project, these are to be taken as accurate.

- **The use of the TK1, and TX2 Nvidia Platform** - The solution will be run and tested on an a platform designed for an embedded system. The assumption that these modules are to be specifically used will contribute greatly to the experimentation, testing and training of the solution.
- **Proof on concept** - The solution will be a proof of concept that reaches a reasonable production standard, therefore it will only identify specifically required key features within the training data.
- **Datasets** - Datasets are to be provided by Thales UK so that the solution may e trained to recognise situations and objects specific to the scenarios given within these datasets.

3.7 Approach

3.7.1 Agile Development

The approach I will be taking to this project, is one of Agile Development. Each component group of my work will be broken down into separate sprints, tackling the specific user stories and requirements that will allow me to solve the problems and limitations that I am aiming to overcome in this project. I will be using regular task lists, milestone markers and log any changes that need to be made from sprint to sprint. In addition to this, I will be keeping to the structure of the project guidelines provided to all honours students by the university, with regular visits to the project clinics when required, as well as regular weekly meetings (or scrums) with my dissertation supervisor, and industrial supervisor, so that my progress can be monitored. Furthermore, with regards to the technical approach, I will be designing my own algorithm(s) and solution structure, as well as incorporating proven methods in computer vision and taking advantage of available libraries and frameworks to make my solution a success.

Chapter 4

Project Plan

4.1 Overview

As I will be using an Agile Development approach to my work, I have constructed a Gannt chart to formalise my planning, each "sprint" is a time period where certain components of the project must be tackled and completed, if they are not completed, these will then be transferred to the next available sprint. There are also milestones included within the chart to represent my personal deadlines as well as given deadlines, dates, presentations and so forth.

4.2 Timeline

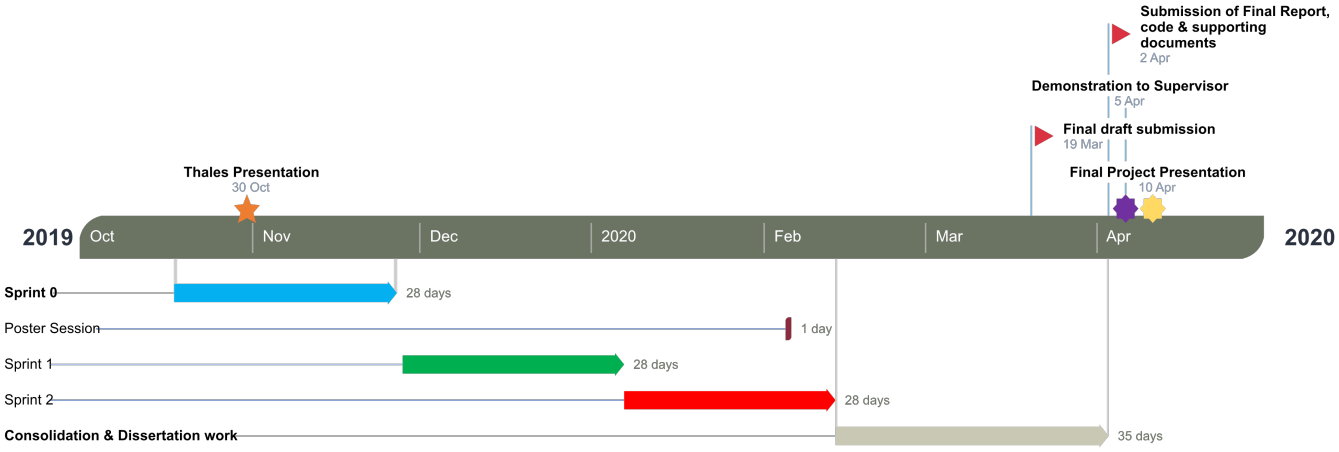

Figure 4.1: Gannt Chart Roadmap

4.3 Deliverables

Name	Delivery Date	Description
First Thales Presentation	30/10/2019	Presentation on solution path & intro.
Poster Presentation	05/02/2020	Poster presentation of project & conversation.
Dissertation Document	2/04/2020	Document recording all project research, & information.
Source Code	2/04/2020	Source code for project solution.
Supporting documents	02/04/2020	Supporting material for entire project.
Demonstration to Supervisor	05/04/2020	Technical Demonstration
Thales Demonstration	TBC	Technical Demonstration

4.4 Intended Result

At the end of Sprint 2 and the final deadline for the submission of my project, I hope to have achieved the requirements necessary to deem my project a success. I hope to deliver a quality proof of concept solution, along with all the supporting material, including this dissertation document, on the intended date of submission. I hope this document to be of high quality, easy to follow and useful to anyone who may read and review it in the future. Towards the end of sprint 2 I hope to demo my finished proof of concept solution to my industrial supervisor at Thales, with a short presentation alongside this, so that my work may be more thoroughly explained.

Chapter 5

References

[1]A. Al-Masri, "How Does Back-Propagation in Artificial Neural Networks Work?", Medium, 2019. [Online]. Available: https://towardsdatascience.com/how-does-back-propagation-in-artificial-neural-networks-work-c7cad873ea7. [Accessed: 17- Oct- 2019].

[2]J. Brownlee, "A Gentle Introduction to the Rectified Linear Unit (ReLU)", Machine Learning Mastery, 2019. [Online]. Available: https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/. [Accessed: 16- Oct- 2019].

[3]J. Brownlee, "Deep Learning Models for Human Activity Recognition", Machine Learning Mastery, 2019. [Online]. Available: https://machinelearningmastery.com/deep-learning-models-for-human-activity-recognition/. [Accessed: 17- Oct- 2019].

[4]"Gradient descent, how neural networks learn — Deep learning, chapter 2", YouTube, 2019. [Online]. Available: https://www.youtube.com/watch?v=IHZwWFHWa-wt=650s. [Accessed: 18- Oct- 2019].

[5]D. Gupta, "Fundamentals of Deep Learning - Activation Functions and their use", Analytics Vidhya, 2017. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/10/fundamentals-deep-learning-activation-functions-when-to-use-them/. [Accessed: 11- Oct- 2019].

[6]A. Karpathy, "ConvNetJS MNIST demo", Cs.stanford.edu, 2019. [Online]. Available: https://cs.stanford.edu/people/karpathy/convnetjs/demo/mnist.html. [Accessed: 18- Oct- 2019].

[7]M. Laboratories, "The gradient descent function", Internal Pointers, 2017. [Online]. Available: https://www.internalpointers.com/post/gradient-descent-function. [Accessed: 17- Oct- 2019].

[8]V. Lepetit and P. Fua, "Keypoint Recognition using Randomized Trees."
Available: https://icwww.epfl.ch/ lepetit/papers/lepetit-pami06.pdf. [Accessed 16 October 2019].

[9]C. Nicholson, "A Beginner's Guide to Neural Networks and Deep Learning", Skymind,

2019. [Online]. Available: https://skymind.ai/wiki/neural-network. [Accessed: 10- Oct- 2019].

[10]M. Nielsen, "Neural Networks and Deep Learning", Neuralnetworksanddeeplearning.com, 2019. [Online]. Available: http://neuralnetworksanddeeplearning.com/chap1.html. [Accessed: 15- Oct- 2019].

[11]S. Saha, "A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way", Medium, 2018. [Online]. Available: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53. [Accessed: 15- Oct- 2019].

[12]J. Shubman, "What exactly does CNN see?", Medium, 2018. [Online]. Available: https://becominghuman.ai/what-exactly-does-cnn-see-4d436d8e6e52. [Accessed: 18- Oct- 2019].

[13]D. Soni, "Supervised vs. Unsupervised Learning", Medium, 2018. [Online]. Available: https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d. [Accessed: 14- Oct- 2019].

[14]T. Strat, Employing Contextual Information in Computer Vision. Menlo Park, California: SRI International, pp. Pages 3-5.

[15]B. Xiao, H. Wu and Y. Wei, "Simple Baselines for Human Pose Estimation and Tracking", arXiv.org, 2018. [Online]. Available: https://arxiv.org/abs/1804.06208. [Accessed: 17- Oct- 2019].

5.1 Figure References

Figure 2.1 - [A]M. Nielsen, "Neural Networks and Deep Learning", Neuralnetworksanddeeplearning.com, 2019. [Online]. Available: http://neuralnetworksanddeeplearning.com/chap1.html. [Accessed: 15- Oct- 2019]

Figure 2.2 - [B]"Neural Networks — OpenCV 2.4.13.7 documentation", Docs.opencv.org, 2014. [Online]. Available: https://docs.opencv.org/2.4/modules/ml/doc/neural(underscore)networks.html. [Accessed: 15- Oct- 2019].

Figure 2.3 - [C]"Logistic function", En.wikipedia.org. [Online]. Available: https://en.wikipedia.org/wiki/Logistic(underscore)function. [Accessed: 18- Oct- 2019].

Figure 2.4 - [D]D. Raschka, "Why is the ReLU function not differentiable at x=0?", Dr. Sebastian Raschka. [Online]. Available: https://sebastianraschka.com/faq/docs/relu-derivative.html. [Accessed: 18- Oct- 2019].

Figure 2.5 - [E]M. Nielsen, "Neural Networks and Deep Learning", Neuralnetworksand-

deeplearning.com, 2019. [Online]. Available: http://neuralnetworksanddeeplearning.com/chap1.html. [Accessed: 15- Oct- 2019]

Figure 2.6 - [F]M. Laboratories, "The gradient descent function", Internal Pointers, 2017. [Online]. Available: https://www.internalpointers.com/post/gradient-descent-function. [Accessed: 17- Oct- 2019]

Figure 2.7 - [G]"Reasearch Gate - A CNN"[Online]. Available: https://www.researchgate.net/figure/A-convolutional-neural-networks-CNN-fig6-321286547. [Accessed: 18- Oct- 2019].

Figure 2.8 - [H]"SuperDataScience", Superdatascience.com, 2018. [Online]. Available: https://www.superdatascience.com/blogs/convolutional-neural-networks-cnn-step-2-max-pooling/. [Accessed: 18- Oct- 2019].

Figure 2.9 - [I]"Semantic Segmentation algorithm is now available in Amazon SageMaker — Amazon Web Services", Amazon Web Services, 2018. [Online]. Available: https://aws.amazon.com/blogs/machine-learning/semantic-segmentation-algorithm-is-now-available-in-amazon-sagemaker/. [Accessed: 18- Oct- 2019].

Figure 2.10 - [J]A. Chen, "Going beyond the bounding box with semantic segmentation", The Gradient, 2018. [Online]. Available: https://thegradient.pub/semantic-segmentation/. [Accessed: 18- Oct- 2019].