Computing Science and Mathematics
University of Stirling

Contextual Understanding and Intent Detection on an Embedded System

Daniel Rooney
2535156

Supervisor: Dr. Deepayan Bhowmik

Dissertation Outline

October 2019

Chapter 1

Introduction

The purpose of this project is to allow pre-existing software, based on an embedded system, to identify objects and their actions and determine their intent via a visual input. This is to be achieved with the use of advanced scene segmentation and intelligent neural networks. By definition, "Intent detection" is the ability to recognise, classify, and confirm an object/persons intent, whether it be through a set of actions, tone of voice or body language. Intent detection and computer vision are therefore inevitably linked. Computer vision applications using intent detection allow users to determine the direction of a situation before it begins, potentially creating safer conditions, increased efficiency, and easier accomplishment of tasks. There are several instances of Computer Vision applications being used in the present, such as face recognition, object tracking and standardised object identification. Applications specific to intent detection are discussed and analysed in the following *state of the art* section.

1.1 Background and Context

The problem to be addressed within the project is the inability of the current software to intuitively determine the behaviour/purpose of an identifiable object. This is something that cannot be achieved without building and identifying contextual elements (whether or not these come from pre-trained CNN with predetermined classes or data sets) around said objects. Potential issues that may arise within the project are primarily linked to the ability of the software to identify separate objects around a subject of interest, and successfully classify/categorise these objects individually or in a group using a pre-trained identifier from existing data sets. Furthermore the potential to be able to construct a possible model around a subject (3D or otherwise), to identify any "hidden objects" that may be essential to determining the behaviour or context of a subject and/or object within its surrounding environment, could prove exceptionally difficult within the projects scope, however feasible this potential method may be.

purus et, vehicula consequat ligula. Phasellus maximus, nisi eu bibendum feugiat, tellus augue malesuada libero, in commodo metus turpis sed lectus. Curabitur pretium varius felis tristique porta. In euismod nisl id velit auctor egestas ut quis ex. Praesent ac tempus tortor, at aliquam eros. Phasellus efficitur erat nulla, ac interdum tortor blandit nec.

1.2 Scope and Objectives

1.2.1 Scope

The scope of this project is to be able to implement a solution that allows the software to analyse and and identify certain contextual cues and behavioural sequences carried out by objects/individuals using activity recognition techniques and semantic scene segmentation, allowing the the software to make a conclusion on the intent of a specific person/object, determine the contextual surroundings of a situation and produce a result that can be categorised as either, no threat, potential threat or direct threat. All of this is to be carried out on an embedded system that can be ran at a relatively low level. The direct results gained from this project allow the operator quickly assess a situation that my be potentially dangerous, and react accordingly, potential preventing harm to themselves or equipment.

1.2.2 Objectives

The objectives that make up the scope of this project can be broken down into three main entities. Firstly, the software must be able to correctly identify and track an object or an individual. Secondly, the software must be able to conclude a contextual understanding of an identified object/person, either by classifying an objects surroundings in relation to that object, or identifying specific activities around an area of interest. Thirdly, the software must be able to recognise a sequence of behaviours or movements that an object or person(s) is undertaking and predict/evaluate the intent of said person within a short amount of time. Recognising activities based on training from previous data-sets, identifying potential objects of threat and comparing them to the context of the situation as well as the object/persons behaviours at that time.

Chapter 2

State-of-The-Art

2.1 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 application and huge advances in what we can achieve with them, it truly is becoming a pillar of the "digital age".

2.1.1 Neural Networks

2.1.2 Hidden Markhov Model

2.1.3 Activity Recognition

2.1.4 Object identification

2.1.5 Supervised & Unsupervised learning

2.1.6 Regression & Classification

2.2 Semantic Scene Segmentation

2.3 Part & Key point Recognition

2.4 Pose & Action Analysis

2.4.1 Microsoft - Human Pose Recognition

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 (enter reference here), 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.

- **The Photogrammetric Context** - This type of context, as expected is defined by the methods of photogrammetry, the photogrammetry process can be broken down into two areas, the exterior orientation of a camera/device - the ability of a device to define its position within a given area, and determine its field of view, relative to where it is, and the interior orientation - which defines the Extrinsic parameters of its actual visual process, for example: the camera calibration, and the association of points in a field of view that can be used to calculate equivalent points in a real world model view. (This is defined as the projection matrix). In summary, it is the context of where the camera/device is in relation to a scene, what it sees and how that translates to the actual physical context.

- **Computational Context** - The computational context is the understanding of all the processes being carried out on a processing level, how these processes relate to each other, and the recollection of the state of theses processes at a given time, this is especially useful for example, when registers are performing tasks, and an interruption is encountered, the ability of the computer to remember it's "position" or its point in a task, what resources it needs to carry out this task and where to begin once it visits the task again.

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.

Hardware - Embedded Systems

The platforms for development to be used are the Nvidia Jetson TK1 and the Nvidia Jetson TK2 for testing on an embedded system. These AI computing devices offer lots of support for a large range of external devices, making them easy to integrate with current systems, as well as being power-efficient, quick and fully supported, with extensive documentation.

Chapter 3

Problem description and analysis

3.1 Problem Description

The existing platform is 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 facing is the inability of the 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.

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 system has 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 system at present has 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 method of notification is a simple box labelling method.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 more so a non-functional requirement, however it is just as important, as it is the area that interacts 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 assumtions as well as additional process information that relates to the formed conclusion.

- **"Must provide an accuracy rating to produce notification"-** Must provide an acccuracy 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 -
textit"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 Use Cases

3.7 Assumptions

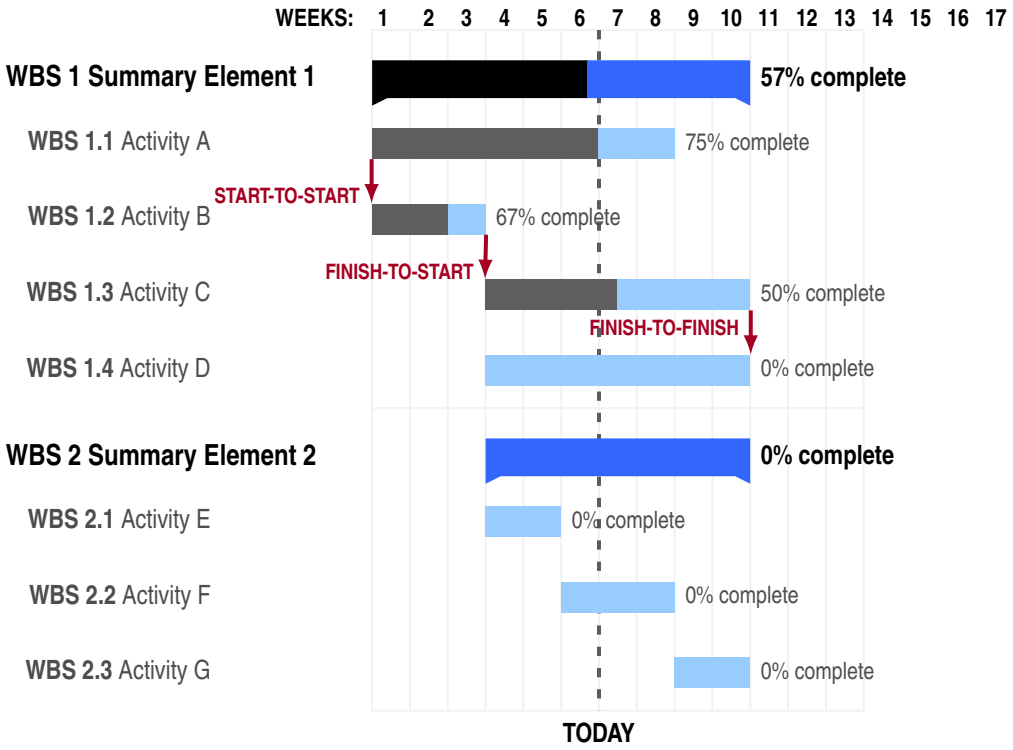
3.8 Approach

Chapter 4

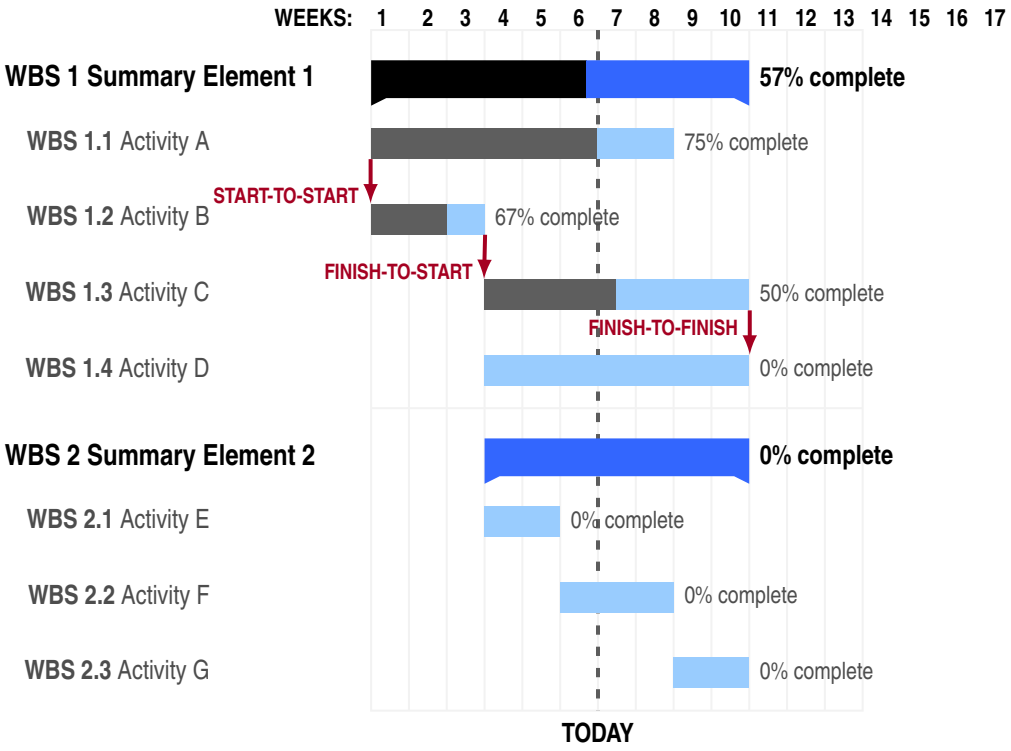
Project Plan

Give an updated plan for the remainder of the project. This should identify the remaining pieces of work (deliverables) and when you plan to complete them. In particular, describe what you will demonstrate when the final report is due.

Autumn Semester 2019



Summer Semester 2019


WEEKS: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
WBS 1 Summary Element 1
57% complete
WBS 1.1 Activity A
75% complete
START-TO-START
WBS 1.2 Activity B
67% complete
FINISH-TO-START
WBS 1.3 Activity C
50% complete
FINISH-TO-FINISH
WBS 1.4 Activity D
0% complete
WBS 2 Summary Element 2
0% complete
WBS 2.1 Activity E
0% complete
WBS 2.2 Activity F
0% complete
WBS 2.3 Activity G
0% complete
TODAY

References