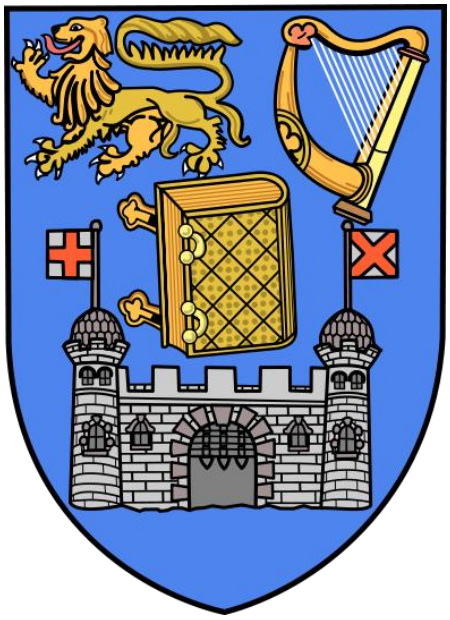
Abstract - Frontal Detection of Backpacks in Surveillance Videos

Ian Beatty-Orr

This dissertation presents a system that can classify an individual as wearing a backpack or not based solely on video footage of the individual from the front. None of the published literature available has presented a reliable solution to this problem that will work in varying illumination conditions. Current video surveillance systems are handicapped by the sheer volume of footage they produce. The aim of this project is to aid operators in sifting through this volume by exploring several solutions that will find and highlight backpacks worn by individuals. An exploration is made of the available literature to examine the approaches adopted to solve similar problems. Following this several designs are developed that search for straps in the upper torso region of an individual. As no suitable test data was available a set is constructed as part of this project to enable testing. It consisted of several videos with subjects, garments, backpacks and locations chosen to attempt to accurately represent the population and scenes encountered by a real camera. Each design was tuned to achieve maximum performance and results obtained against this test set. The design process was iterative with later approaches addressing weakness that became apparent in earlier approaches. The results generated are promising with the best approach achieving an accuracy of 79.5%. This demonstrates that computer vision can be used to detect backpacks within a scene.

Frontal Detection of Backpacks in Surveillance Videos



Author: Ian Beatty-Orr

Supervisor: Dr Kenneth Dawson-Howe

A dissertation submitted to the University of Dublin, Trinity College, in partial fulfilment of the requirements for the degree of M.A.I. (St.), April 2014

Declaration

I, Ian Beatty-Orr, declare that the following dissertation, except where otherwise stated, is entirely my own work; that it has not previously been submitted as an exercise for a degree, either in Trinity College Dublin, or in any other University; and that the library may lend or copy it or any part thereof on request.

Signature:

Date: *Wednesday 2nd April 2014*

Acknowledgments

Firstly I would like to thank my supervisor Kenneth Dawson-Howe for the time and effort he has put into helping guide me during this project. Without his input the system presented in the following pages would not have evolved to be as successful as it is. My parents and sister also deserve significant thanks for their continued support and encouragement throughout the year. They have put up with proof reading this document time and time again. Last but not least I would also like to extend my thanks to both the Electronic and/or Computer Engineering M.A.I. year of 2014 and my two roommates Michael and Tony. They ensured that this year (as well as the previous four) was an enjoyable experience and kept my work life balance in check.

Summary

This dissertation documents the development and testing of several approaches to the task of identifying backpacks on individuals from video surveillance footage. In particular this report looks at detecting a backpack from a front view only as previous work has concentrated mainly on side views.

The first step in this project was an exploration of the available literature to document previous attempts at backpack detection. Since there are very few published papers available on the subject the initial research focuses on general detection techniques and solution used to detect items similar to backpacks. The research then focuses in on the area of detecting backpacks from side views as constitutes most of the published material. These approaches are analysed to see if any of them can be applied to the process of detecting backpacks from the front. However in the end it was concluded that these approaches were not suitable for the process of detecting from the front only hence any solutions developed would have to be done so explicitly.

Before the process of developing new solutions could begin a way of verifying and benchmarking the performance of them needed to be devised. The literature review and additional intensive searching has un-covered no suitable sources of test data. All available test databases were too low in numbers to be accurately representative of a population or did not provide the full frontal views required for this project. An ideal test set will represent the conditions that will be encountered by the average camera. This would consist of numerous combinations of garments, scenes and backpacks to try and test the system as comprehensively as possible.

Hence a new custom test data set was developed for the purposes of this project. This involved the selection of as many people as possible to try and attain as high a diversity of gender and body shapes to ensure the tests were representative of the general population. A variety of locations were then scouted out to ensure the system could cope well in both indoor and outdoor conditions under a variety of lighting conditions including twilight. Different combinations of garments and backpack straps were used to try and simulate the variety that would be encountered in real life. The aim is to make this data set publishable so that other researches can use it to verify the results of this project and any subsequent publications as well as benchmarking their own systems against the test data. It should be noted that the ability to benchmark a system against third party data is enormously beneficial to the field of computer vision. It eliminates bias between the test data and system under test, an advantage not available to this project.

Overall there were six main approaches that developed through iterative improvements made to each other. These can roughly be divided into two groups the first based on colour space analysis and the second based on gradient (change in image intensity spatially) analysis. All of them concentrated on locating the backpacks straps within the upper torso region of detected individuals.

The first approach was colour space analysis applied to the whole image. The colours within the image were clustered into approximately 3 groups (this number could be varied). These groups were analysed statistically for properties such as their location, height to width ratio and symmetry. This initial approach proved to be a total failure as illumination changes in the image tended to

outweigh the colour changes causing in-correct clustering. This was solved in the fifth[i] approach by concentrating on the fact that backpack straps give the strongest colour change response in the horizontal direction. Hence by clustering along rows instead of the whole image the colour change would outweigh the illumination change and other irregularities. When combined with statistical checks of the symmetry and strap width variance over the rows this proved to be a very successful method. It was applied to the process of searching for only single straps in the sixth approach.

The gradient based approaches all concentrated on trying to detect the edges of the strap as this usually produced a strong gradient response. Like the colour space analysis this produced the strongest response along the x-axis hence only the gradient response in this direction was used. Approach two (the first gradient based approach) only took a localised snapshot of edge magnitude and direction for an arbitrary number of rows. This proved unsuccessful as there was no check for continuity between each of these rows. Approach three solved this by taking into account the spatial positioning and orientation[ii] of edges and searched for parallel pairs. This approach had a high recall rate usually detecting backpacks when they were present however it had a high false detection rate as it also detected backpacks that were not present. This problem was addressed by approach four which analysed a colour histogram of the potential strap regions and compared them to the colour histogram for the non-strap regions to eliminate these false detections.

Of these solution approach four and five represented the pinnacle of development for the two groups and achieved accuracies of 70.5% and 79.5% respectively. All approaches produced several parameters that had to be tuned to produce optimal performance of the system. This involved running against the test data several times for different values of one parameter to find the optimal. This process was then repeated for the remaining parameters and once they had been optimised the entire process was repeated iteratively until the results stabilised.

At the end of this project a representative test set that can be used the verify he performance of frontal backpack detectors has been created. In addition two very good approaches at detecting backpacks using colour analysis and gradient analysis respectively have been developed. These show that computer vision can be used to detect backpacks on an individual in a variety of conditions. These include twilight conditions in low level outdoor scenes, brightly lit indoor and outdoor scenes. Striped garments have been checked against with good levels of success. Moderate success has also been encountered when looking for single strap backpacks. A list of improvements that could be introduced to the systems to further improve their accuracy has been listed in the future work section of this dissertation.

---

[i] Due to total failure of approach one colour space analysis was abandoned and not returned to until after the three gradient based approaches had been developed, hence the irregularity in jumping from approach one to five.

[ii] Spatial Orientation refers to the angle the edge makes along its length with the x-axis as opposed to gradient orientation which indicates the angle the gradient at a specific pixel makes with the x-axis.

Table of Contents

1 Introduction 1
1.1 Motivation 1
1.2 Project Goals 2
1.3 Outline 2
1.4 Terminology 4
2 Literature Review 5
2.1 Introduction 5
2.2 Detection of People 5
2.3 Foreground Segmentation 7
2.4 Object Detection and Classification 7
2.5 Side on Backpack Detection 9
2.6 Frontal Backpack Detection 12
2.7 Summary 16
3 Test Data, Ground Truth and Testing Process 17
3.1 Introduction 17
3.2 Requirements of Test Data 17
3.3 Currently Available Test Data 18
3.4 Construction of Test Data Set 20
3.5 Annotation of Test Data with Ground Truth 22
3.6 Testing Process and Result Generation 23
3.6.1 Measures of Accuracy 23
3.6.2 Testing Process 24
3.6.3 Tuning Parameters 24
3.7 Summary 25
4 Design Requirements, Pre and Post Processing 26
4.1 Introduction 26
4.2 Requirements 26
4.3 Development Environment 27
4.4 Pre-Processing: Isolation of Upper Torso Region 28
4.4.1 Detection and Tracking of People 29
4.4.2 Foreground Segmentation 29
4.4.3 Foreground Restoration 31
4.4.4 Isolation of Upper Torso Region 31
4.5 Post Processing: Combining Individual Frames 32
4.5.1 GUI Window 34
4.6 Summary 35

5 Approach One: Colour Space Clustering and Statistical Classification 36
5.1 Design Overview 36
5.2 Colour Clustering 36
5.3 Statistical Analysis 37
5.4 Results 38
5.5 Evaluation 39
5.6 Summary 39
6 Approach Two: Edge Gradient and Orientation Analysis 40
6.1 Design Overview 40
6.2 Edge Detection 41
6.2.1 Gradient Analysis 41
6.2.2 Non-Maxima Suppression 42
6.3 Orientation Analysis 43
6.4 Results 43
6.4.1 Successes 45
6.4.2 Failures 46
6.5 Evaluation 46
6.6 Summary 46
7 Approach Three: Parallel Edge Analysis 47
7.1 Design Overview 47
7.2 Parallel Contour Extraction 48
7.3 Symmetry Analysis 49
7.4 Results 49
7.5 Evaluation 50
7.6 Summary 50
8 Approach Four: Parallel Edge and Colour Space Analysis 51
8.1 Design Overview 51
8.2 Colour Histogram Analysis 52
8.3 Comparison 52
8.4 Results 53
8.5 Evaluation 55
8.5.1 Successes 55
8.5.2 Failures 57
8.6 Summary 59
9 Approach Five: Row by Row Colour Space Clustering 60
9.1 Design Overview 60
9.2 Clustering Individual Rows 61
9.3 Statistical Analysis of Individual Rows 62

9.4 Statisitcal Analysis of All Rows 62
9.5 Additional Statistical Checks 63
9.6 Results 63
9.6.1 Successes and Failures 64
9.7 Evaluation 66
9.8 Summary 67
10 Approach Six: Single Strap Detection 69
10.1 Design Overview 69
10.2 Differences from Approach Five 69
10.3 Results 70
10.3.1 Successes and Failures 71
10.4 Evaluation 71
10.5 Summary 71
11 Conclusion and Future Work 72
11.1 Introduction 72
11.2 Comparison of Approaches 72
11.3 Conclusion 73
11.4 Future Work 75
A. Methods A-1
A.1 Haar Face Detector A-1
A.2 Colour Spaces A-2
A.3 K-Means++ A-4
A.4 Detailed Annotation System A-4
A.5 Attached CD A-5
B. Results B-2
B.1 Parameter Tuning Approach Four B-2
B.2 Parameter Tuning Approach Five B-3
B.3 Best Results for Each Approach B-4
B.4 PR Curve Approach Four B-5
B.5 P-R Curve Approach Five B-5
C. References C-1

1 Introduction

1.1 Motivation

Modern surveillance systems arose from a need; security personnel were not able to keep an eye on everything at once. Cameras were installed to provide extra sets of eyes both in real time and the ability to review events at a later date. The sheer volume of footage produced by modern systems prevents their full potential from being realised. The London Underground and Heathrow Airport each have over 5,000 individual cameras [2]. It has been estimated that the U.K. as a whole has between 4 and 5.9 million CCTV cameras [3]. Unless an operator is actually viewing a camera at a particular point in time it may as well not be there. Even when the cameras record what they see, there will still be an immense volume of footage to sift through after an incident. In 2011 52 analysts spent 14 days viewing 5,000 hours of footage collected after a post baseball game riot in Vancouver [4]. The analysts spent a significant portion of time classifying people as wearing baseball caps, jackets backpacks and so on. This was necessary to help link a view of an individual committing a crime to a view of their face in another scene for the purposes of court proceedings.

Current surveillance systems are un-intelligent and un-able to recognise what they are seeing. The ideal solution would be a system that can recognise and respond to what it sees without requiring an operator [2]. However the available technology is not yet at a stage where this can be implemented. One step that could be taken towards this ideal would be to only present the operator with footage of interest. This would be achieved by classifying scenes as having people present or absent and classifying those people based on visually distinguishable features.

There are solutions available that will automatically detect people and vehicles[5]. In addition work has already been done on detecting backpacks when viewed from the side [6]. This project is going to focus on detecting a backpack when viewed from the front. This will complement current research as very little has been conducted in this area so far. This will be a challenging task as only the two straps will be visible from the front. The visibility and contrast of these straps will vary by a large amount depending upon the underlying garment, scene and illumination conditions encountered.

Backpacks have been involved in two recent terrorist events, the London and Boston bombings both of which involved explosives being carried in backpacks on individuals in public locations [4, 7]. Hence it would be of interest to the security world if they could be tracked reliably on individuals as they move around. Another useful application of backpack tracking is to supplement existing systems that can detect abandoned and removed objects in a scene. However if a person was going to plant a bomb for instance as in the Boston Marathon it would be easy to fool these systems by placing the bomb out of camera shot. Detection by inference involves designing the system to note the presence of a backpack on a person as they walk through a scene. Therefore we are able to notice if a person is wearing a backpack walking into a specific room and not wearing a backpack when they walk out of a room a while later.

1.2 Project Goals

There are two main goals of this project:

- To create a challenging and publishable dataset representative of the general population to enable testing of frontal backpack detection systems.
- Design and evaluate several new techniques for detecting backpack straps within an image.

The reasoning behind the first goal is due to the limited amount of test data currently available for video surveillance projects. They do not contain enough scenes of people wearing a backpack walking towards the camera for this project. To enable adequate testing of the methods developed for the second goal of the project will require the creation of a custom dataset. Ideally this will be one that can be made publically available for other researchers to use. In the field of computer vision it is standard practice to release the test data when publishing. This enables other researchers to verify results and compare the performance of their own systems.

To give respectable results the test data needs to recreate natural occurrences as much as possible and avoid being scoped towards one scenario. Hence time and effort needs to be invested into producing scenes that are likely to occur in real world situations. This will entail a variety of different test candidates, garments and backpacks to ensure test data that reliably represents real world conditions is created. By making the data publically available for other researchers to use will make future research into the area of frontal backpack detection significantly easier and enable benchmarking of systems against third party data.

The second goal of this project is to develop several explicit methods that look for the unique features of backpack straps within a video sequence. The techniques will be designed from the ground up to keep them simple and hopefully achieve optimal performance. Ideally these methods should be able to correctly classify a backpack as being present or not in only a few frames. This classification should not be adversely affected by:

- changing lighting conditions
- individuals walking at a slight angle towards the camera
- Different combinations of colours for the strap and underlying garment

1.3 Outline

The remainder of this dissertation is structured as follows:

Literature Review – The main focus of this chapter is investigation of the currently available literature on how to detect objects within video sequences. Starting off with general object detection the focus of the chapter narrows down to previous work on detecting backpacks from a side view. This chapter does not serve to give the reader an introduction into the field of computer vision and a basic knowledge of computer vision techniques is assumed. If the reader is not familiar with certain techniques explanations can be found in the appendix chapters.

Test Data, Ground Truth and Testing Process – The third chapter catalogues publically available test data that was considered for use in this project. It then explains why it was concluded that none of these were adequate for the needs of this dissertation and the only alternative was to create a new dataset. The chapter then looks at the methodology of obtaining the test data in such a way as to make it representative of the conditions likely to be encountered by the system. It finishes of with a brief look at how this data was annotated with ground truth values to enable live evaluation of results.

Design Requirements, Pre and Post Processing – The aim of this chapter is to give the reader the information they require to understand the subsequent chapters on each individual solution. This enables the reader to proceed straight to the solution they are interested in without having to read all of the previous solutions. The chapter starts with a look at the requirements that need to be satisfied for a computer vision based backpack detector to be considered successful. It then explains the common pre-processing methods applied before any of the solutions are used. It finishes off with a quick explanation of how results are obtained.

Approach One: Colour Space Clustering and Statistical Classification – Chapter five will begin with a high level design overview of the final method employed by solution one. Following this will be a discussion of the theory of the methods used and the design decisions as to why these methods were chosen. It will then present solution ones results explaining the successes and failure cases. The results will then be evaluated relative to the requirements that were laid down during the design phase.

Approach Two: Edge Gradient and Orientation Analysis – Identical in structure to chapter five.

Approach Three: Parallel Edge Analysis – Identical in structure to chapter five.

Approach Four: Parallel Edge and Colour Space Analysis – Identical in structure to chapter five.

Approach Five: Row by Row Colour Space Clustering – Identical in structure to chapter five.

Approach Six: Single Strap Detection – Identical in structure to chapter five.

Conclusions and Future Work – The final chapter briefly compare the results obtained from all of the different approaches and states the conclusions of the project in light of the goals set in the section above. It also presents the areas of the project that the author feels warrant further evaluation and investigation if the project were to be continued.

1.4 Terminology

True Positive (TN) – Backpack detected when present

False Positive (FP) – Backpack detected when not present

True Negative (TN) – Backpack not detected when not present

False Negative (FN) – Backpack not detected when present

Dataset – Collection of videos which the system can be run against

Training Dataset - Dataset used for training and tuning the system

Testing Dataset – Dataset used exclusively for testing the system

CCTV – Closed Circuit Television

EKF – Extend Kalman Filter

IMM – Interacting Multiple Model Filter

UTR – Upper Torso Region

MCC – Matthew Correlation Co-Efficient

2 Literature Review

2.1 Introduction

This chapter explores and categorizes state of the art literature relating to the utilisation of computer vision for the detection and classification of objects in surveillance footage. It does not give the basic background to common computer vision operations nor does it expand on the algorithms used. Further information on these two components can be found in the appendix chapter.

There are vast amounts of literature available on the topics of detecting people within a scene and removing the background. These two vital pre-processing steps are reviewed within sections 2.2 and 2.3 respectively. Section 2.4 examines current work on the detection and classification of objects similar to backpacks with the view to finding suitable techniques that can be applied to the process of detecting backpacks. Significantly less literature is available regarding backpack detection. The majority is based upon symmetry analysis of the human silhouette when viewed from the side and this area is examined in section 2.5. A gap is present in the published literature when it comes to detecting backpacks from the front, an area that silhouette based detectors do not work in.

2.2 Detection of People

In order to classify an individual as wearing a backpack or not we first need to reliably detect that a person is present in a scene and locate them. This field has been well studied and various solutions are available.

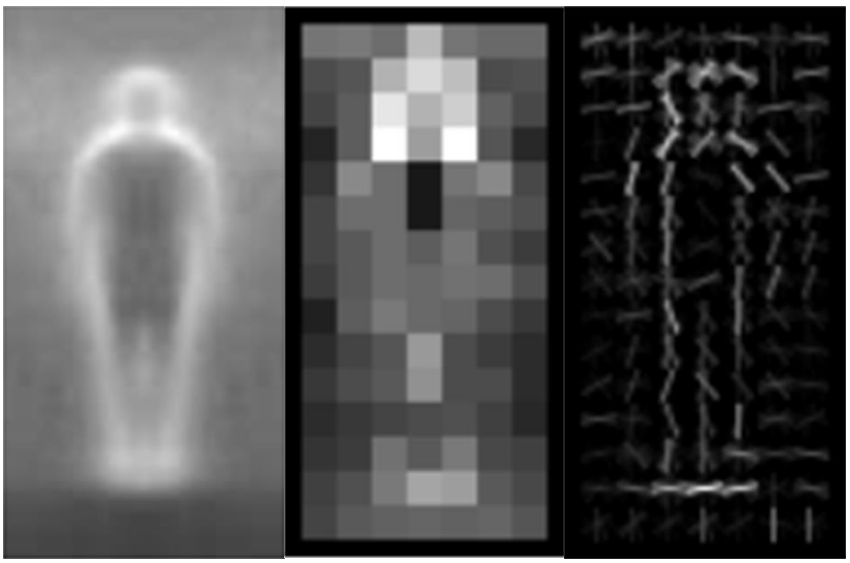


Figure 1a: Gradient space representation of a human created by compositing several hundred gradient images.
Figure 1b: N*N blocks with intensity representing their weight, bright regions being important.
Figure 1c: Graphical representation of the orientations present in each N*N block (divided into orientation 9 bins), intensity represents the number of pixels that fall in that bin by their magnitude.
Three images taken from [5]

Histograms of Oriented Gradients (HOG) as outlined in [5] is a technique that searches for the full human body. Templates of how the average human being appears in gradient space have been compiled as shown in *fig 1a*. The rectangular region around the person in the template is then split into *N*M* blocks of size P. For each of these regions a histogram is created for the orientation of all the pixels within a block weighted by their magnitude. The histogram for each of these blocks is then weighted by its position within the rectangle as more importance is given to blocks at the edge of the person as indicated in *fig 1b*. This process is then repeated for several positions of the rectangle around the image as indicated in *fig 1c.* The histograms obtained for these regions are then fed into a Support Vector Machine (SVM) using the templates histograms as ground truth. If the SVM gives a strong enough response a bounding rectangle is given to indicate the location of the person.

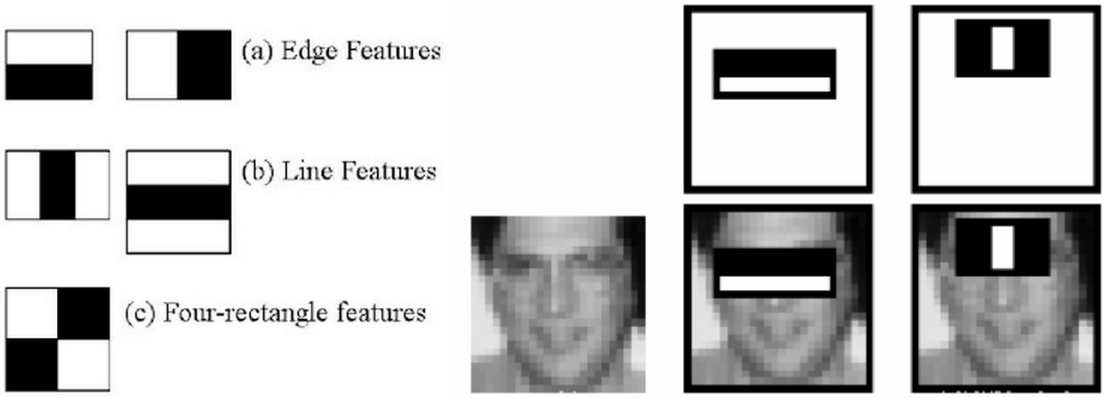


Figure 2a: Haar Wavelet Shapes used in Viola and Jones Face detector, [8].
Figure 2b: Haar wavelets being applied to an image of a face,[8].

Another approach for detecting people is to detect faces as they approach the camera. This solution has the drawback that it only works when the person is facing the camera, a drawback not applicable to detecting backpacks from the front. Detecting faces is advantageous in a crowded scene where parts of the body would be occluded behind others causing a full body detector such as the HOG to fail. [9] discusses a method for detecting faces based upon intensity differences within images based upon regions defined by Haar Wavelets as shown in *fig 2b*. The reasoning is that if one of these regions is placed over a feature, such as the eye and cheek, the overall intensity of the eye region will evaluate as darker than that of the cheek. Individually a classification such as this is incredibly weak and almost guaranteed to fail. However by combining hundreds of these classifications in stages using a process known as AdaBoost the combined classification can be quite strong. The system developed in Viola and Jones [9] checked 6,000 features over 38 stages. As with the HOG the classifier needs to be trained with significant amounts of ground truth, [10] mentioned that 5,000 positive samples were used and a similar number of negative samples. The detector can be trained to detect whole faces, only face silhouettes, eyes, ears and other specific features. An extended discussion and explanation of Haar Wavelet based detection can be found in *appendix A1*.

A slightly less accurate but computationally superior approach to the Haar wavelet detector is Local Binary Patterns (LBP) as discussed in [11]. The idea of checking many weak features and combining

them into stages is the same as used with the Haar. However instead of classifying features by Haar wavelets LBP is used. LBP classifies a pixel as being either a 1 or a 0 depending on its greyscale representation relative to all of the pixels around it.

2.3 Foreground Segmentation

Figure 3: Original image on left, a mask showing the detected foreground regions in white on the right with shadows in grey.

Another necessary step in the process of detecting backpacks on a person is to remove the background from the scene. This ensures that later classification steps do not end up evaluating pixels that do not belong to a person or backpack. The simplest approach is background subtraction where a single image is taken of only the background scene with no moving objects present. This image is then subtracted from future frames leaving foreground objects behind as shown in *Figure 3*.

Needless to say this model is overly simplistic and has been expanded upon, one approach is a Gaussian Mixture Model (GMM) as proposed by [12]. Here a model of the background is built up over several frames. For each pixel in the image several Gaussian distributions are created, one for each of the common values that the pixel can have. For all subsequent frames each pixel is compared against the Gaussians to see if the value of the pixel fits one of them. If it fits one of the models it is considered a background pixel otherwise it is considered a foreground pixel.

One of the main issues encountered when using a GMM or any other form of background subtraction is that shadows cast by foreground objects can result in those region being incorrectly classified as foreground regions. Hence [13] developed a method where the GMM is adaptive and can differentiate pixels not fitting the model due to being a foreground object or merely a shadow cast by a foreground object.

[14] proposes a version of GMM that can cope with variable lighting conditions. The algorithm was developed to respond to the problem of background subtraction in an atrium lit by a skylight. A robust foreground subtraction method will be required that can work in a variety of conditions including indoor and outdoor with changing lighting environments.

2.4 Object Detection and Classification

Due to the absence of literature available on detecting backpacks from a frontal view it is advisable to examine several techniques that have been used in other more general object detectors.

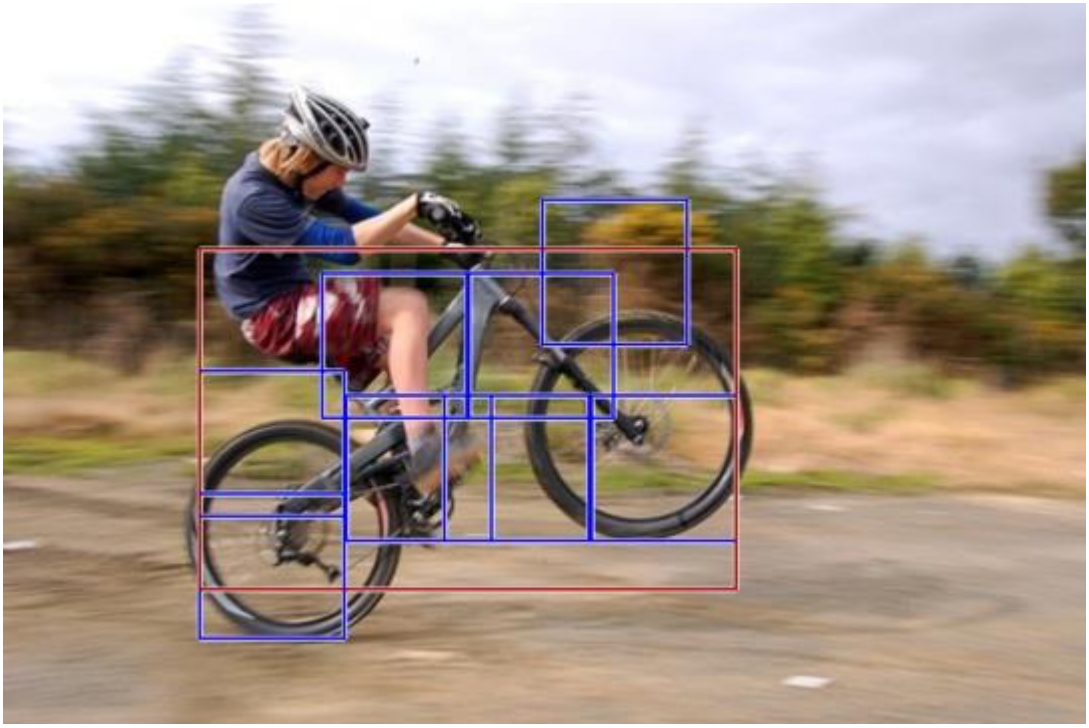


Figure 4: Deformable Part Model being applied to a bicycle, red box is the root filter and blue boxes are the feature detectors

Several computer vision based systems make use of a deformable part model as described by [15]. This finds an overall root match for the whole object within a frame. It will then try to find matches for distinguishable features of the object as shown in *fig. 4*. If enough of these features are found the correct distances from each other the item will be determined as a match. The advantage of using this approach is that the detection criteria can be made very sensitive with high detection rates ensuring all instances of an object are caught. It is the low probability of the root detector and several features being falsely detected that ensures the accuracy of the system. This option also copes quite well with objects that will change appearance as they change orientation relative to the camera. Take for instance a bicycle at 45 degrees and at 90 degrees relative to the camera, the same features can be seen in both. In the 45 degree case they will be closer together but still visible.

[16] have augmented this with an Extended Kalman Filter (EKF) to enable the tracking of bicycles as they move through a 3D environment. This was extended further with the use of an Interacting Multiple Model Filter (IMM) in. The IMM filter is better suited to scenarios where the direction of an object is changing rapidly. However this comes with additional computation cost as the IMM is a multi-model filter versus the simpler single mode EKF.

[16] [17] implemented a system that makes use of either an Extend Kalman Filter (EKF) or an Interacting Multiple Model Filter (IMM) to track the location of bicyclists relative to a moving vehicle. This makes use of a HOG detector to detect bicycles within the scene as well as AdaBoost to improve the classification accuracy and speed.

One of the more interesting applications was present in [18] with the interesting problem of how to differentiate between bicycles and motorbikes. They adapted the strategy of searching for specific combinations of Gabor wavelets in the Y direction only as these were produced by the peddling motions of a cyclist. Their overall system made use of a deformable part model with a HOG to recognise the individual parts followed by a SVM classifier to differentiate between people, bicycles and motorbikes. This study was able to correctly detect 99.6 of bicycle/motorbike objects and 100% of pedestrian objects. However when an attempt was made at differentiating bicycles from motorbikes the correct detection rate was 40.9% for the former and 61.7% for the latter.

In my view these papers show that a good strategy to adopt when detecting general objects is to make use of a Haar or HOG wavelet detector with a deformable part model using a weak root filter and stronger features to ensure false positives are minimised.

2.5 Side on Backpack Detection

If we take a look at pre-existing work into the field of detecting backpacks we can see that there has been some investigation into the area. However all of these methods require the main body of the backpack to be visible to the observing camera and protruding from the wearers' silhouette. Hence all of these methods only work in the case of side on viewing.

One of the earliest and best known examples is Backpack [19] which is based upon the W4 person detector[20]. Many approaches have been based on this pioneering 1999 system which extracts silhouettes from the background and then analyses their symmetry, segmenting out asymmetrical portions. The periodicity of each asymmetrical segment is compared to that of the overall silhouette over a human gait cycle [21]. If the periodicity of a segment is very close to the overall body. It is classified as part of the body (for instance an arm or leg swinging back and forth). When the periodicity is significantly different the segment is classified as a carried object.

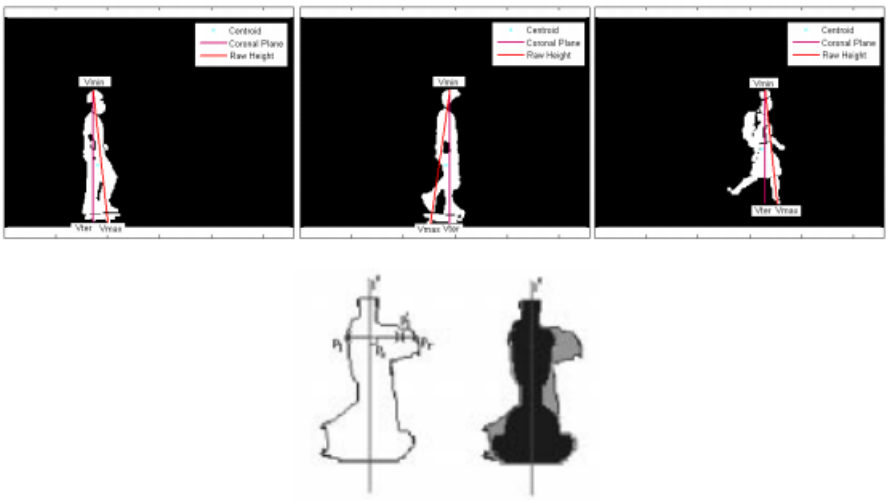


Figure 5: Symmetry Analysis of a Silhouette showing determination of the centreline (top) and asymmetrical regions (bottom) [HARITAOGLU '99]

The backpack method was expanded upon by [22] to work on thermal images in a night time environment using more advanced gait curves to work out the asymmetry as shown in *fig. 5*. The advantage of this system was that it could be tuned to re-move the backpack from the human silhouette and classify the person based upon their gait curve. When running the system against the CASIA night database [23] a successful classification rate of up to 95% was obtained under controlled conditions.

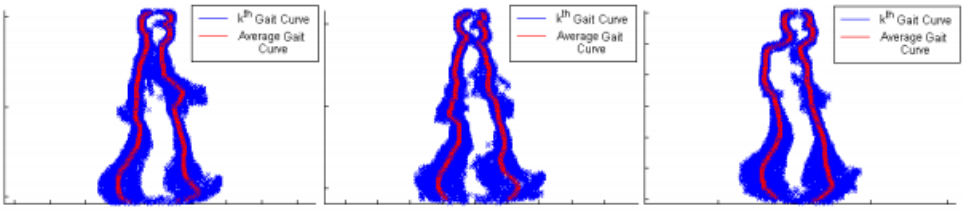


Figure 6: Gait curves for three different subjects. The mean gait curve is marked in red. [CHIRAZ '02]

Gait analysis was expanded upon by Chiraz BenAbdelkader [24] to determine if a person was carrying a backpack or not based upon motion analysis, *fig. 6*. This method centres on the fact that a person's gait cycle will change if they are encumbered with heavy objects such as a backpack. Their stride will shorten and the frequency their feet make contract with the ground will increase. When run against 41 sequences the system achieved a successful detection rate of 85% and a false alarm rate of 12%. This was supplemented in [25] by the examination of Gabor wavelets to improve gait detection.

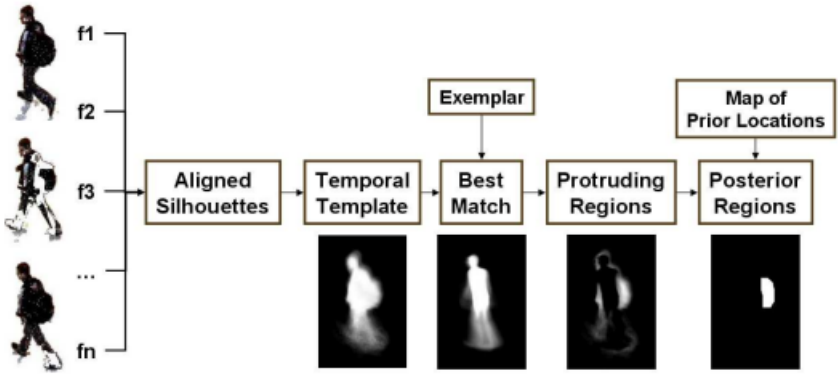


Figure 7: Method of backpack detection where a temporal template is constructed of the walking person and compared to an exemplar of an un-encumbered person to find protruding regions. [DAMEN '08].

Another silhouette based approach was presented in [26], shown in *fig. 7.* Instead of analysing the symmetry straight off silhouettes were concatenated over a human gait cycle. This was compared to a 3D Maya model which was constructed from exemplar templates created by observing eight people walking without any carried objects in eight directions. This exemplar model is then scaled and rotated to find the best fit relative to the concatenated silhouette image and the two are then compared to find protruding regions. This addressed some perceived weaknesses that caused failures in Backpack:

- Human gait periodicity changes due to the weight of the carried object.

- The displacement of the central axis of symmetry due to the pixel mass of the protruding carried object.
- Other errors related to the calculation method that may distort the gate and periodicity leading to incorrect classification of asymmetric regions.

	Precision	**Recall**
Thresholding	39.8 %	49.4 %
MRF - Prior	50.5 %	55.4 %

When run against the PETS2006 database the above results were obtained. Analysis revealed that false positive detections were caused by protruding body parts and pieces of clothing, extreme body proportions as well as incorrect and noisy template matches and duplicate matches. False Negatives were caused by bags with little or no protrusion, dragged bags separate from the individual, carried objects not segmented from background, two protruding regions merging into one, swinging objects, object being carried between legs as well as noisy and incorrect templates and shadows being miss-detected.

An extended version of this system [27] had periodicity analysis incorporated into the system to try and improve the detection rate. In addition the system was now run over both the PETS2006 data set and the LEEDS 2009. The latter aimed to address perceived weaknesses of the former which focused mainly on indoor locations. However there does not seem to have been a significant change to the results presented.

Figure 8: Interaction between two people where one individual runs away afterwards. [28]

I will finish this subsection with a brief explanation of two novel security based systems. First a theft detection system [28] that uses colour histogram analysis of detected people. When two tracked individuals come together in a scene the colour histograms of each individual before and after interaction is compared. If change in both histograms indicates an item was exchanged between the individuals it is investigated further. The actions of both people are then fed into a finite state machine. If one of them runs away the system considers the exchange as a theft and raises an alarm *fig. 8*. If both walk away it considers it as a normal acceptable exchange of a bag.

The second [29] makes use of neural networks in monitoring archaeological sites in Egypt. This system aims to identify when people are carrying certain classes of objects such as pic-axes and shovels that can be used for activities such as grave robbing. Hence it helps to reduce grave crime in Egypt.

The main issue with all of the systems presented so far is that they require the bag to be visible directly to the observing camera. Hence large amounts of errors have been caused by protruding cloths or body parts or objects being too small or flat to stand out from the silhouette. In particular all of these systems require the person to be viewed side on to the camera. Most of them will fail to cope with a person walking directly towards the camera. Hence frontal backpack detection will be my next area of investigation.

2.6 Frontal Backpack Detection

There is only one recent paper that has partially addressed the issue of detecting a backpack when viewing an individual walking directly towards or away from the camera. This was published after this project implementation was underway [30]. It examines the possibility of detecting single strap bags and backpacks on people from multiple angles in crowded environments.

The system starts out by tracking human heads using LBP and HOG avoiding the need for the full human body to be visible. The upper torso region of the person is then estimated based upon head position. This estimate is refined by applying colour histograms and K-Means clustering to the temporary upper torso region to remove any portion of the lower body included.

Figure 9: Head detected with yellow box at top, initial and refined estimates in blue and red boxes respectfully. Thresholded upper torso region shown on right.

Adaptive thresholding is used to segment the upper torso region into two regions as shown in *fig. 9*. This paper works on the assumption that the darker region is the strap and the brighter region is the underlying garment. The system works for two cases:

- Single Strap Sling Bags
- Double Strap Backpacks

Single straps are located by two means, first geometrically and secondly using Canny edge detection. The geometric method analyses the darker regions of the thresholded image to look for blobs that

have a very high length to width ratio indicating a strap as shown in *fig. 9*. However this can often fail so two parallel edges indicating a strap are searched for again by running a Hough detector over an edge image created using Canny. This is done in two passes with the geometric search first and the Canny/Hough approach only used if the first method fails to detect a strap. A person is classified as wearing a sling style bag if one strap is found from any angle.

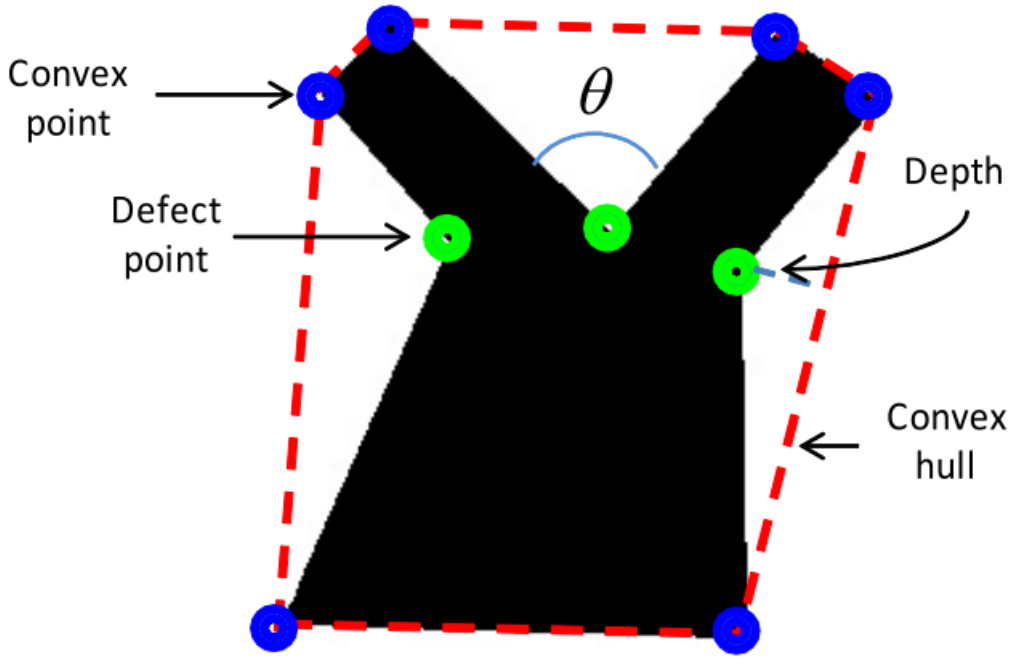


Figure 10: Silhouette of Backpack when viewed from behind. Note the three V-shaped concavities indicated by the green circles.

When searching for a backpack what is visible will be different depending on whether the individual is being viewing from the front, rear or side:

- From the front the single strap method is modified slightly to classify the individual as wearing a backpack if it detects two straps.
- From the rear the geometric method is extended by also finding the convex hull of the dark region. A typical backpack will produce three V-shaped concavities, one at the top and one on each side as shown in *fig. 10*. Only two of these need be detected for a backpack classification to be applied.
- From the side only either the left or right V-shaped concavity need be detected.

Figure 11: Several Failures cases labelled a-f respectfully

This system fails due to several cases:

- Texture differences between the underlying garment and the backpack straps such as checked shirts as indicated in *fig. 11c*.
- Hair occluding the straps and or merging with the underlying garment is another major issue that causes failures of the system as shown in *fig. 11d & e*.
- Asymmetric strap sizes as indicated in *fig. 11f.*

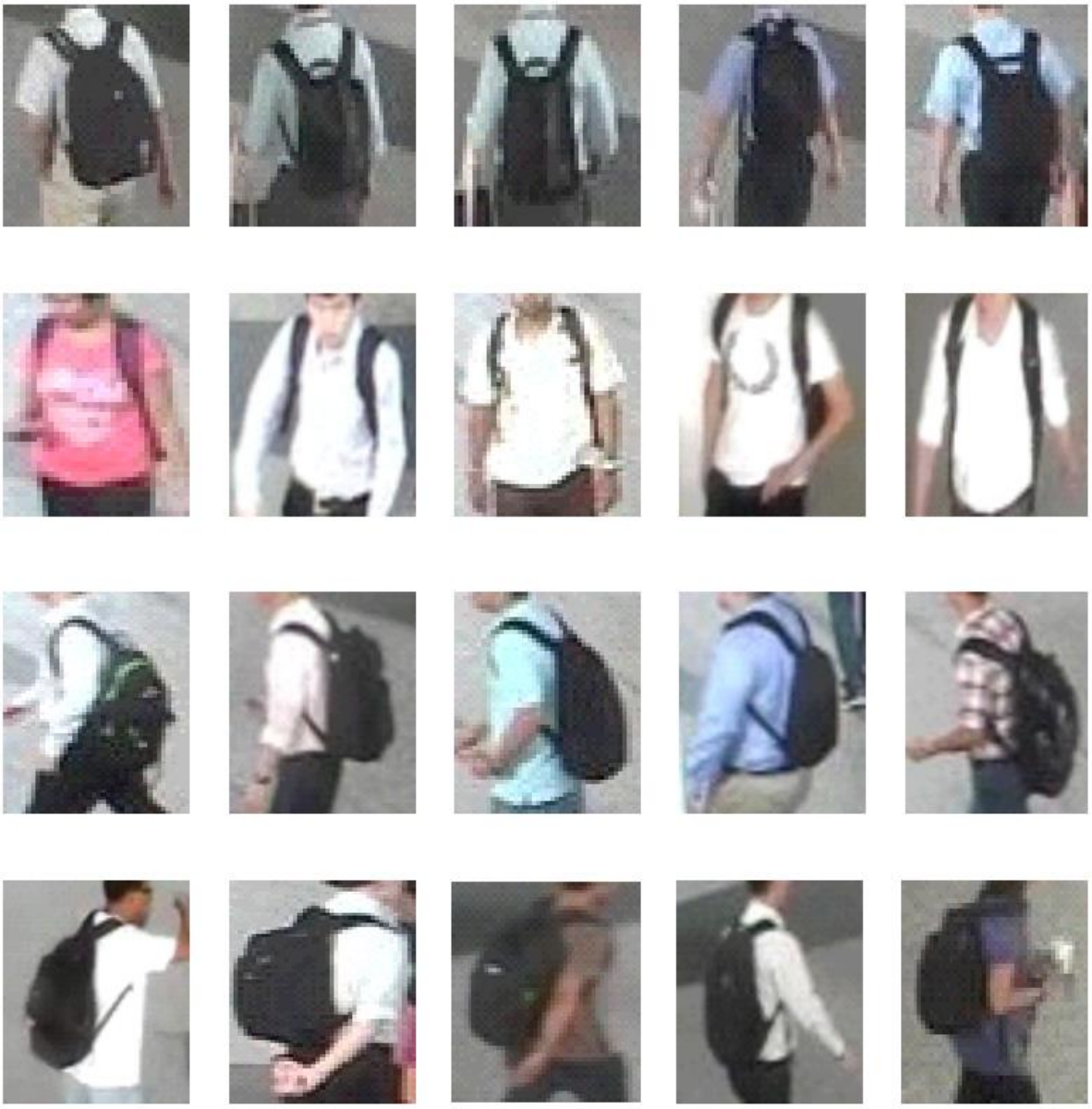


Figure 12: Examples of Backpack used as test data.

Upon evaluation it appears to that the test data used by this study is biased. It appears that most of the footage was collected from an airport in a warm climate. Hence most of the garments worn are t-shirts and shirts with a light colour. A sample of the test data used can be seen in *fig. 12*. The publication does not present anything darker than the brown t-shirt in the middle of the bottom row. Hence all of the straps have a high level of contrast relative to the underlying garment. This is in contrast to the conditions the author observed while obtaining test footage in Ireland on a February morning where the majority of clothing worn was dark coloured coats and jackets giving very low levels of contrast relative to the straps. Hence I believe this publication's results would not be so high if the study were repeated in our climate during the winter or using the test data set provided later in this paper.

Additionally all of the test data was obtained from a well-lit indoor terminal building. As will be presented later in this paper systems can often encounter difficulties when poorer more variable lighting conditions are encountered outdoors.

2.7 Summary

A video surveillance system designed to detect backpacks on people needs to be able to cope with any angle a person may be visible from. As can be seen a lot of work has been done on locating and tracking multiple individuals as they walk through a scene. Significant work has also been done on segmenting foreground objects from the background. There are already systems available that detect backpacks in side views based on their protrusion from the human silhouette. However a surveillance system will need to cope with cases where the backpack itself is not visible such as when the person is walking directly towards a camera. While one paper has laid out basic work in this area it needs to be expanded to increase robustness and reliability, particularly for the more difficult illumination conditions encountered outdoors. This area is a gap in the available knowledge that this thesis will attempt to answer.

3 Test Data, Ground Truth and Testing Process

3.1 Introduction

When publishing in the field of computer vision it is standard practice to include a sample dataset as supporting evidence of the system's performance. The purpose of this is twofold; first it provides support for the presented results and enables them to be verified. Secondly it allows other researchers to benchmark the performance of their own systems against the outcome of this project. As there was not enough existing test data for the purposes of this project new data had to be created. This will contribute to currently available data and save other researchers the time and effort of having to create their own data and provide them with the advantage of having a test set created by a third party available.

This chapter begins by looking at the requirements for a test set to ensure it adequately represents real world conditions and test the system. Section 3.3 explores the currently available data sets and states why they do not satisfy the requirements of this project. Section 3.4 looks at the construction of a custom data set for the purposes of this project. The task of annotating this data set with ground truth is discussed in section 3.5

3.2 Requirements of Test Data

The ideal test data set will have been produced by a reputable external source and be publically available [31]. This ensures that the test set and system were developed separately and avoids the introduction of bias. Being publically available enables the test set to be used as a benchmarking piece of data. This enables easy comparison between systems as they will have results based upon the same dataset.

A good test set will also span a broad range of conditions to ensure it tests as many of the situations the system is likely to encounter as possible. In addition to obtaining a wide range of cases a good test set will also feature more examples of commonly occurring conditions than un-common ones. This means obtained results will reflect accurately on how the system will perform in the real world. Often the best test data has been recorded naturally without participants having any knowledge the recording was taking place or any attempt made to alter the conditions present within the scene.

Here is a list of the some of the different requirements needed to ensure that test data is broad enough in its scope to cover as many different aspects as possible:

- Having a wide variety of locations is necessary to ensure that specific features present in the background of one scene do not affect the results. It also increases the chance of a weakness within the system to certain kinds of background being highlighted. For instance in one of the locations used there was a tendency for table legs to occasionally be confused as people.

- Using video data that has both indoor and outdoor scenes is extremely beneficial. A system that performs well under controlled lighting conditions indoors may fail in the variable illumination conditions encountered outdoors.
- When detecting backpacks on individuals we want to have as many different combinations of backpacks and underlying garments as possible. It will be fairly easy to detect backpacks with high contrasting straps relative to underlying garments and near impossible for straps with no contrast difference. However the real test will be how the system handles straps with a low contrast that are still visible to the human eye. In addition it will be interesting to see how the system can handle different combinations of texture and patterns on the garments. Vertical stripes on a garment may cause confusion and trick the system into classifying them as containing a backpack.
- We would like the test data to contain people being viewed from several angles. A person wearing a backpack walking towards the camera will display two perfectly symmetrical straps. However if they are walking at a slight angle relative to the camera the strap on the far side will become narrower than the near side strap from the cameras perspective. This could confuse the system and will be beneficial to test against.
- **Occlusion** – In crowds the straps can become occluded by other people, in addition people may carry objects which cover the straps or people may have two bags on at once.
- **Odd Carrying Arrangements** – Often people will wear a backpack with only one strap over a shoulder. This could confuse the system as it will only see one strap.
- **Open and Closed Jackets** – Garments with zips are likely to cause extra edges to appear in the middle of a garment which may confuse an observing system.

3.3 Currently Available Test Data

As was stated in the requirement section using test data that is already publically available and created by a third party will increase the reputability of the results and save time. There are several publically available data bases that may fulfil this role. Below are listed the data sets that came the closest to fulfilling the requirements of this project.

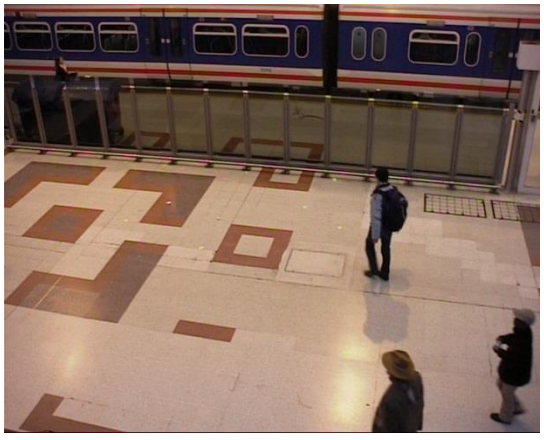


Figure 13: PETS 2006 database

- **PETS 2006** – This is a dataset based in a large train station in England aimed at examining left luggage. However the camera angles presented are not directly in the line of pedestrian flow presenting very few frontal views. Hence this dataset is more suited to a system searching for a backpack from the side. There are also very few scenes that focus on backpacks as the data set focuses on left luggage in general such as suitcases and hold all's.

- **PETS 2007** – This was a collection of test samples taken from a busy airport however many of the people appearing on screen appeared at a resolution that was too low nor where they on screen for long enough.

- **PETS 2009** – This was an artificial scene of several people walking around a crossroad in Birmingham University. It came closer than any of the other databases to providing useful test data as several people were wearing backpacks and walking directly towards the camera. However the cameras were positioned very far away from the crossroad resulting in a low pixel width of individuals. Hence the number of pixels representing the straps was too low for evaluation. There was also very little variety present in the combinations of garments and backpacks encountered. Mainly dark colours on dark colours and hard for even a human observer to see.

Figure 14: PETS 2009 image showing low contrast backpack on the individual on the right at too low a resolution.

- **CASIA Gait Database Dataset C** – This was different to the previous datasets as it contained individuals recorded using thermal imaging cameras. Unfortunately this dataset is not publically available and only the silhouettes can be downloaded. This would be useful for analysing side on based backpack detection as discussed in the literature review but not for the purposes of this project.

After searching as many test databases as possible using sources such as Engineering Village, Web of Science, Google Scholar and IEEE no suitable test data could be found. Therefore the only solution available was to construct a custom one for the purposes of this project.

3.4 Construction of Test Data Set

This thesis is looking at the task of detecting backpacks when viewed from the front only. It is not looking at cases where the person is walking side on to the camera or away from the camera. Nor is it trying to improve on person location or background subtraction techniques any more than has been achieved in other publications. To avoid creating situations that may cause a failure of these methods and to concentrate solely on the challenge of detecting backpack straps, the test sequences were kept simple. These all consisted of one individual walking either directly towards the camera or at a slight angle to the camera. For each person the clip was taken three times, once with both straps of the backpack on, once with only one strap on and finally without the backpack on. When testing the system against the data, either only the single strap or double strap positive sequences would be used to preserve the balance between positives and negatives. This ensured that the number of sequences that should return a negative was equal to the number of sequences that should return a positive. It is important to keep an equal number of both scenarios to avoid masking a system that has a tendency to give false positives or vice versa.

Before filming could begin ethical approval had to be obtained from the school of computer science and statistics. Once this was obtained the test sequences subjects were chosen from students of Trinity College Dublin. There was a mixture of ethnicity and gender to ensure the system was tested on a diverse range of people and to avoid the system becoming biased to certain body types, for instance tall males. As wide a range of garment and backpack combinations as possible was tried in order to increase the chance of a combination that caused the system trouble being detected.

Figure 15: Examples images from the test data created for use in this project.

Video Number			Flags			Garment	Backpack
Positive	**Single Strap**	**Negative**	**P**	**S**	**A**		
714	-	713	M1	I1		Black jacket	Black Straps
717	-	716	M1	I1		Red/white polo shirt horizontal stripes	Black Straps
719	-	718	M1	I1		Dark Brown jacket	Black Straps
737	-	733	M2	O1		White/grey pull over horizontal stripes	Black Straps
738	-	733	M2	O1		White/grey pull over horizontal stripes	Single Black Strap
736	-	735	M2	O1		Black Jacket with open zip above white/grey pull over horizontal stripes	Single Black Strap
740	-	739	M2	O1		White/grey pull over horizontal stripes	Black Straps
746	745	744	F1	I2		Olive Green shirt with red scarf	Black Straps
753	749	*754*	M3	I1		White Jumper	Black Straps
756	757	755	M4	I1		Navy Blue Jumper	Black Straps
760	768	*765*	M5	I2		Dark Green Jacket	Black Straps
761	766	*764*	M6	I2		Grey Hoodie	Black Straps

769	771	773	M6	I2		Grey Hoodie with red scarf	Black Straps
770	772	774	F1	I2		Olive Green Shirt	Black Straps
792	793	794	M7	O2		Dark Grey Hoodie	Black Straps
796	795	797	M3	O2		Grey Hoodie w text	Black Straps
798	800	799	M1	O2		Red/grey hoodie	Black Straps
801	801	803	M7	O3		Dark Grey Hoodie	Black Straps
804	805	806	M1	O3		Red/grey hoodie	Black Straps
815	816	817	F2	O4		Navy hoodie	Maroon Straps
849	850	851	M1	I3		Blue t-shirt	Black/Maroon Straps
856	857	858	M1	I3		Red/white horizontal striped polo shirt with black and grey scarf	Black/Maroon Straps

Table 1: Details of Videos contained in Test Data

Images from some of the test sequences are included in *fig. 14*Figure 15 as well as more detailed information in *Table 1*. In total there were 22 cases for a total of 44 videos when testing the double strap case. 16 of these cases also had an additional single strap case that could be used when evaluating the ability of the system to detect single straps. There were seven male test subjects and two female test subjects. Three different indoor scenes were used and four different outdoor scenes. These sequences were all recorded with a 14.1 megapixel camera and converted into video at a resolution of 1280 by 720.

3.5 Annotation of Test Data with Ground Truth

The purpose of annotated data is to provide a reference to evaluate the performance of the system against. The annotated data provides information on the actual presence of a backpack within a system. There were two annotation systems used in this project, the first was a more complex system that provided information on the presence of a backpack within each frame as well as the location of the backpack within the frame. It required the author to go through each video recorded and mark the presence of a backpack and straps for each frame as well as draw bounding boxes. The information provided by this annotation system was fond to be greater than necessary to generate results. In addition the complexity of the annotation system made it time consuming to annotate each frame of every video and also increased the potential for annotation errors due to operator fatigue. Additional information is available on this annotation system in *appendix A3*.

To reduce the time needed to annotate videos and increase reliability a second annotation scheme was created that annotated whole videos rather than individual frames. This system gave each video one of four classifications:

1. Backpack Present
2. Backpack not Present
3. Occluded Backpack Present

4. Fake Backpack Present

The third classification was used if there was a backpack within the sequence that was hidden under another item such as a scarf or there was a single strap bag crossing over the backpack straps. The fourth classification was used if there was an item present within the scene that would confuse the backpack detection method into thinking there may be a backpack such as a scarf.

3.6 Testing Process and Result Generation

Presenting results for a binary classification system such as this one is not as straightforward as it initially seems. Depending upon the requirements of the system either false positives or false negatives may be considered more severely than the other. Hence this subsection will examine several measures of accuracy that take both of these errors into account. It will then detail how the system was tuned and tested against the test set to maximise performance using these measures. The actual results are presented for the four systems individually in subsequent chapters before being compared in chapter 11.

3.6.1 Measures of Accuracy

The binary classification system presented in this thesis will give two possible results: either a backpack is present or it is not. There may in fact be a backpack in the scene or there may not. This leads to four possible outcomes when evaluating binary classification results as shown in the confusion matrix below.

	Predicted Positives	Predicted Negatives
Actual Positives	**TP**	**FN**
Actual Negatives	**FP**	**TN**

Figure 16: Confusion Matrix

There are many basic measures of the performance of binary classification systems, most of them are simple combinations of the above four parameters. This report will concentrate on accuracy, precision and recall however there are others available such as sensitivity and specificity as used in medical research.

- **Accuracy** – This measures how many predictions were correct relative to the overall number of tests: $Acc = \frac{TP+TN}{TP+FP+TN+FN}$
- **Precision –** The number of predicted backpacks that are correct: $P = \frac{TP}{TP+FP}$
- **Recall –** The number of actual backpacks that were detected: R = $\frac{TP}{TP+FN}$

The latter two can be graphed as a P-R curve as shown in *appendix B4***.** The precision and recall values of different approaches can be compared using such a curve as various parameters are tuned

to try and achieve maximal performance. The goal is to try and get points in the top right of the curve as neither a high precision or recall value is beneficial if the other is too low.

However for the purposes of tuning the system it will be easier if there is only one value representing the accuracy of the system to compare to. The basic accuracy measure detailed above does not take into account the ratio between false positives and false negatives. It will give the same accuracy value to a result set with an even distribution of both as it does to one with twice as many false positives and no false negatives.

A more complex measure of accuracy that takes both precision and recall into account is the F1 score: $F1 = 2 * \frac{P*R}{P+R} = \frac{2*TP}{2*TP+FP+FN}$. The F1 score gives a value of 1.0 for a perfect system and a score of 0.0 for the worst imaginable system.

However the F1 score still does not take into account the True Negative rate. A much better measure is the Matthews Correlation Co-efficient which is a balanced measure of binary classification systems that was introduced in 1975 [32]. It returns a value of 1 for perfect classification, 0 for classification equal to random and -1 for classification that is worse than random classification. $MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

3.6.2 Testing Process

To generate the result values for TP, TN, FP and FN from which all other scores were generated the approach under test was run against all of the video sequences in the test data. This is done in a linear fashion with all of the 44 videos being played one after the other as due to the processing requirements no more than one video can be run at a time. For each video the tests results are recorded and compiled with the results from all of the other videos. One run took about 15 minutes to complete on the lab machines being used which had 3.4 GHz Sandybridge quad core processors and 16GB of RAM.

3.6.3 Tuning Parameters

Each of the approaches had several parameters that could be tuned to improve the performance of the system. Changing the setting of one parameter will change the optimal setting of other parameters. Hence the system had to be tuned one parameter at a time in an iterative process. Once an optimal value had been chosen for each parameter in the first round the process had to be repeated with all parameters set to the new values. Ideally this was continued until all of the parameters settled at optimal values. The parameters were all tuned relative to the Matthew Correlation Co-efficient with the aim of achieving a value as close to 1 possible.

For each run, 5 values were tried for each parameter which took around 1.5 hours. A script was used that allowed 6 different parameters to be checked in one night on one machine. To expedite the process the simulation was run simultaneously on several computers in the lab. This process could only be completed at night as the PCs had to be available for students to use during the day time.

3.7 Summary

This chapter has detailed the important steps required to ensure that the system is adequately tested. The creation of test data was a major step in this process as there was no pre-existing test data that could be used to ensure the system functioned as intended. Hence the author had to invest significant time to create a test set that was as representative as real data as possible. This has been achieved and additionally candidates have agreed to make the test data publishable which will provide a useful resource for other researchers. They will be able to benchmark the performance of their systems against the test set and compare their results. In addition this means that the results of this system can be verified by other researchers. This test set has been created to closely represent expected real life conditions as much as possible.

The annotation of this test data with ground truth has also been discussed. This is an important step as it ensures that the system can automatically be tested in a very efficient manner. It should be noted that the reliability of the annotation system was ensured by the development of a second simpler ground truth annotation system.

The testing process and measures of accuracy have also been detailed in the section. A consistent testing process is necessary to ensure that the accuracy of the system remains consistent and does not vary from test to test. Hence results can be reliably compared.

4 Design Requirements, Pre and Post Processing

4.1 Introduction

The purpose of this chapter is to lay out the requirements of the design as well as the common pre and post processing steps. The idea is that this chapter will give the reader enough knowledge on the shared components of the system to jump straight to the approach they are interested in reading about. It is highly recommended that the reader reads the pre-processing steps in Section 4.4

The section will start off by taking a look at the features that would be required in a backpack detection module if implemented in a video surveillance system. The development environment will then be briefly discussed in section 4.3. Section 4.4 explains in detail the common pre-processing steps required for all subsequent approaches. This is followed by the common post-processing steps in Section 4.5.

4.2 Requirements

The requirements of a real world backpack detection module will be extensive as it has to cope with un-controlled conditions and a wide variety of people and backpacks. Many of these requirements such as the ability to track a person in multiple cameras have already been solved in other studies [33]. Hence I will only concentrate on the specific requirements to detecting backpacks in particular the elements that are not solved or have not been solved well at this point in time.

- **Detecting and Tracking an Individual** – To successfully track a backpack throughout a scene relies upon the successful tracking of the individual wearing it. There are several algorithms already developed that extract the human body from a scene and track it such as HOG [5].
- **Isolating the Upper Torso Region** – Backpacks are secured to the wearer by straps that will only be present in the upper torso region. To ensure accurate classification the system should evaluate only this area and cut out unnecessary information from the rest of the body and background.
- **Detecting a Backpack from All Angles** – In an airport, large station or any other open space people will be walking in several directions. A backpack will appear differently to a camera depending upon the angle. For instance the front of a person displays only two straps, the rear of a person displays a large blob with two short sections of strap visible on top. From the side only the rear of the backpack is visible protruding from the person's body and in-between these angles the system will have to cope with a combination of two of these views. There is already significant work available on detecting a backpack from the side [6] and one paper covering detection from the front and rear [30]. This thesis will concentrate efforts on filling the current gap in knowledge by detecting a backpack more robustly from the front. This will be a useful angle to concentrate on, as cameras positioned above a doorway that people are streaming into will only receive a front view. This is also likely the

first angle a camera will capture an individual at when they enter a building with many cameras such as an airport terminal.

- **Detecting a Backpack in Different Illumination Conditions** – Many systems such as those indoors in large airports as noted in [30] will have the benefit of controlled lighting. However cameras positioned at the outdoor entrances to buildings and around public squares will not have this benefit. Cameras positioned outdoors will have to cope with different lighting changes as clouds obscure the sun on some days and rain causes interference on other days. In addition they will have to handle dawn and dusk conditions. There are even studies examining the use of thermal vision cameras for observation at night time [22].
- **Detecting a Backpack with Low Contrast** – Depending upon the colour of the backpack and the material worn by the person there can be great differences in the contrast of material. Predictably distinguishing backpacks and straps with a high contrast relative the underlying garment is going to be an easy task. However other backpacks and items of clothing may have a low level of contrast. Hence the system needs to be able to cope with these low contrasting backpacks or it will fail to detect people in certain scenarios.
- **Detecting a Backpack when the underlying Garment is patterned** – If the garment underneath the backpack has a distinctive or repetitive pattern such as checkers or stripes, it may confuse the system into thinking there are straps present when there not. Other types of pattern may also mask straps and make them harder for the system to pick out.
- **Unusual Wearing Conditions** – While backpacks are usually worn with two straps many people only wear one strap.
- **Occlusion** – If the system were to be used in large crowds there is the danger that people will start to occlude each other blocking the camera's view of the person. In addition depending on what way the person holds another item such as when carrying a coat or second bag it may block parts of the backpack that were used for recognition. Hence a successful real world system will need to cope with occlusion.

4.3 Development Environment

This backpack detection system has been developed using the language C++ with the OpenCV computer vision library version 2.4.6. It was developed with two versions of Visual Studio, 2010 and 2012 both on the Windows 7 operating system. The working application has been tested on two different machines. First a Dell Inspiron 15R SE laptop with an Intel Core i7-3612QM 2.1 GHz quad core CPU and 8 GB of RAM. Secondly a Dell OptiPlex 9020 desktop with an Intel Core i7-4770 3.4 GHz quad core CPU and 16 GB of RAM. Needless to say significantly higher run speeds were encountered on the desktop.

4.4 Pre-Processing: Isolation of Upper Torso Region

All of the approaches used to detect a backpack upon a person required the same initial steps to isolate the upper torso region for further analysis. This involved tracking an individual through a scene, background subtracting and then location of the upper torso region. This section first gives an overview of the design with the following subsections giving a more detailed description of each step.

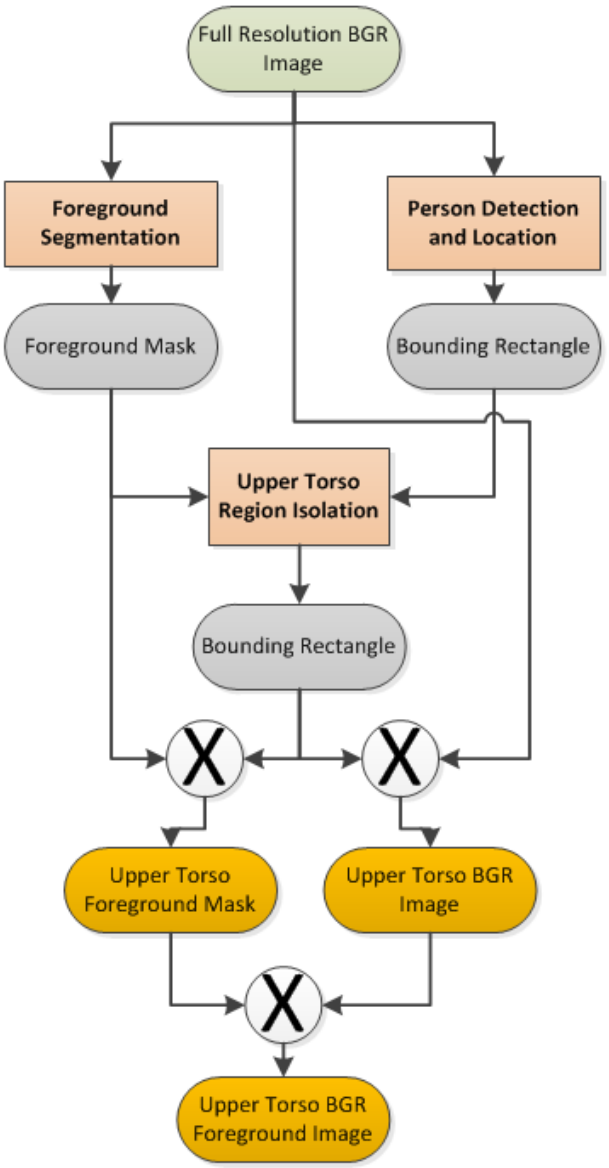


Figure 17: Pre-processing flow chart

1. **Individual Tracking** – The full resolution input image was searched for human bodies. The full human body was easier to detect particularly when the pixel width of the individual was low.
2. **Background Subtraction** – This system is only interested in detecting backpacks on moving individuals. Hence we eliminate the background and stationary objects from consideration creating a foreground mask for the whole image. Detected shadows are not included in the foreground mask.
3. **Foreground Mask Smoothing** – Opening and Closing operations are applied to the foreground mask to eliminate noise. This noise is caused by individual pixels being in-correctly classified.
4. **Gap Filling** – All holes within the mask are filled along with horizontal gaps between any two elements of the foreground mask. Such holes and gaps are usually due to background subtraction failure and we want them to be considered as foreground. Vertical gaps are left unfilled as they tend to destroy the boundary between foreground and background objects in the neck region of the image.
5. **Upper Torso Isolation** – The top of the individuals head and both shoulders are located using the background subtracted image. From these positions a region of interest is statistically generated centred upon the upper torso region. This region is then used for further analysis in subsequent methods.

4.4.1 Detection and Tracking of People

The aim of this step was to take the un-altered full resolution input image and locate any individuals present within the scene. As detailed in previous sections this needs to be able to work in crowded locations where occlusion would be a real concern and at varying distances. The initial design choice was to use a Histogram Oriented Gradient Detector (HOG) to scan the scene for any potential human like objects and return a bounding box around the location. This produced a number of false positives which were dealt with by re-scanning within the box using a Haar detector as described in *appendix A1* to look for faces.

However this did not work as intended as the HOG relies on the full outline of the person being visible hence this method does not work when the person is occluded or comes very near the camera and their legs disappear from the scene.

To try and get around this the Haar was used on its own however it tended to generate a lot of false positives. The solution to this problem was to run two classifiers, one looking for the overall outline of the face and the second looking for a pair of eyes within the face. Both were required to signify a face and this reduced the number of false positives. At the other end of the scale when the person is more distant from the camera the number of pixels representing their face is quite low. This causes the Haar detector to fail. Hence the range of either the HOG or Haar detector on its own is not appropriate for the task. The range where both worked together was even smaller and too constrained for a real world application.

Hence the ideal solution adopted was a combination of both two methods with either one of them detecting a person triggering detection. This had the advantage of the long range of the HOG being able to detect individuals in the background while the Haar could detect individuals very close to the camera and individuals within a crowd. The Haar was also replaced with the slightly less accurate but much less computationally intensive LBP detector to speed up the approach. This approach minimised the number of failed detections at the expense of letting a higher number of false positive detections through. This change was reverted as additional run time was not significant while using the Haar. During the process of testing and tuning parameters the run time of the system was found to be too great. Analysis showed that the Haar was the main performance drag on the system and this was removed from the system. The HOG function fine on its own for test clips with single people in them such as the ones used here however the Haar should be re-enabled if the system is to be used in crowded situations.

4.4.2 Foreground Segmentation

The aim of this step was to take the full resolution un-altered input image and segment it into the foreground region and the background region. As this system was designed with a stationary camera in mind, this problem was suitable for solving with a background model based method. This involves initialising a background model over several frames of just the background. Foreground objects transitioning through the scene can then be determined by comparing the current frame to this model.

Background subtraction is a challenging step as the method needs to cope with changing illumination in the scene, slight movement of the camera and other effects. These effects will result in a change in pixel value which must not be mistakenly classified as foreground. However when a new foreground object enters the scene, the method needs to successfully detect all of these pixels that are part of this object, while separately classifying any shadow introduced by the object as part of the existing background.

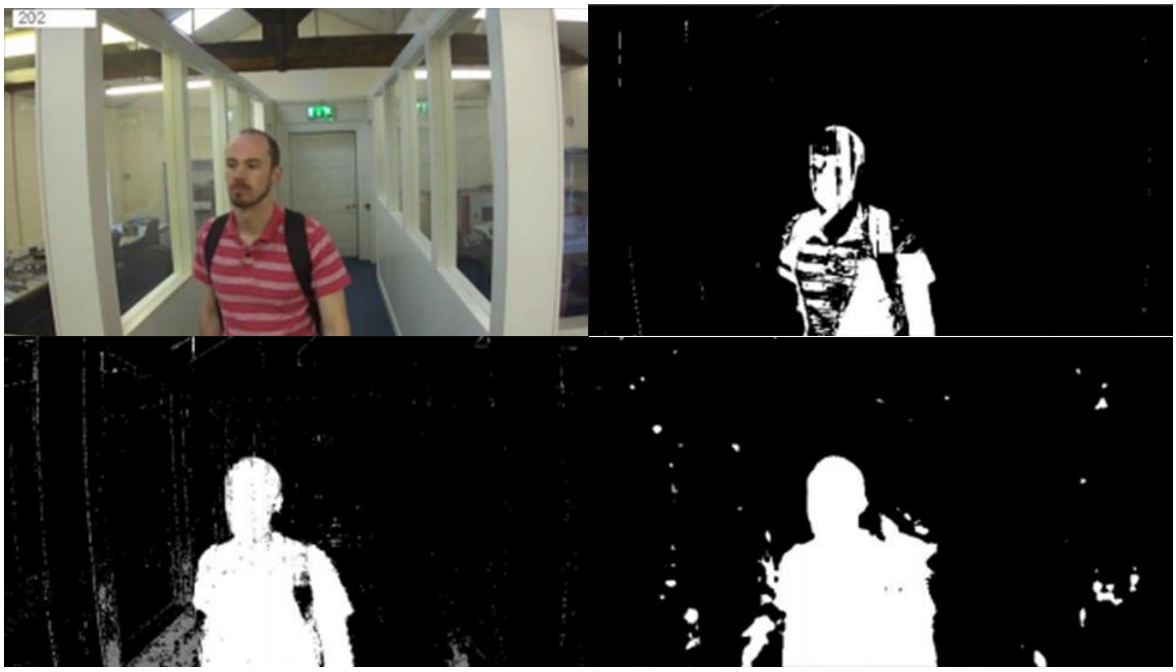


Figure 18: Clockwise from top left: original input image, background subtraction using method of; Kaewtrakulpong '02, Zivkovic '04 and Godbehere '12

The results of three different background subtractors based upon [12], [13] and [14] are shown in *fig. 18* respectively. The 2004 method was selected for continuation as while the 2012 worked better in variable lighting conditions, it often detected too many pixels and had very blurred edges as can be seen there is significant destruction of the neck region.

However due to the challenging nature of image segmentation there would inevitably be misclassified pixels, holes within the object and noise like pixels over most of the image. The noise like pixels can be removed from within the foreground region by performing a closing operation **appendix** where the region is first dilated by a certain amount of iterations and then eroded by the same number of iterations. This was followed by a smaller closing operation, the inverse of an opening, to remove noise pixels from the background of the image. These were conducted in this order to preserve the extent of the foreground region which was strengthened by the opening operation. Opening and closing operations could only deal with the removal of small noise like misclassification as the number of iterations has to be kept low to avoid destroying the edge information. Solving larger misclassifications is dealt with in the next subsection.

4.4.3 Foreground Restoration

As can be seen in *fig. 19a* the foreground image can have holes caused by pixels being incorrectly classified as is evident by the right strap. This can be seen extended a step further where the head visible in *fig. 19r* has been detected as several blobs. This can be solved by tracing contours around the detected foreground regions. These can then be analysed and any region completely enclosed by another can be eliminated removing holes from the foreground region.

To solve large gaps that are not completely enclosed such as the blue regions in *fig. 19* the following method was used. A convex hull was drawn around each contour region, indicated by the red line in *fig. 19*. The original foreground region was subtracted from this to leave the region shown in white in the image below. All of the pixels within this white region were analysed and if they lay horizontally between two the edges of the foreground region rather than the convex hull they were added to the foreground region, the blue region in *fig. 19*. This was only done in the horizontal direction as it was found performing this operation in the vertical direction destroyed the edges of the neck region.

Figure 19: On the left is the original image with filled in regions shown in blue. On the right we can see how the disjoint regions around the head have been combined.

As can be seen in this image the head has been badly segmented and is represented by three blobs. This problem was solved by analysing all foreground regions within the image and if they were close enough together relative to the width and height of the larger contour they were joined as can be seen on the right. This would be run subsequent to hole filling but before the previous method hence why we can see the blue regions applied to this contour.

4.4.4 Isolation of Upper Torso Region

As we now have the bounding rectangle for either the full body or the head location and a background subtracted image we can isolate the upper torso region for further analysis. As further analysis and classification will be based upon this step it is important that this region is accurately located. It also needs to include the top of the shoulders as backpack straps will extend over them.

Figure 20: Locating the top of the shoulders using the foreground mask.

The location method achieves this by first locating the top of the head and shoulders. An initial bounding rectangle of the upper torso region is devised by taking the top half of the bounding rectangle for the person produced by the HOG. This is shown as the green box overlaying half of the red box produced by the HOG. Starting at the centre top pixel of this bounding rectangle every pixel along the magenta line in *fig. 20* is checked until we hit the first pixel in the foreground region. This is taken to be the top of the head. From this height we iterate down ten other lines positioned roughly where the shoulders should be, indicated in yellow and cyan. The median height at which these ten lines find the foreground regions is taken to be the top of the shoulders.

As long the height of the head is above the shoulders a new estimate of upper torso position can be made. This starts from just above the shoulder estimate and has a height scaled from the original height of the person. The width of this new upper torso region is scaled from the width of the person.

4.5 Post Processing: Combining Individual Frames

The four approaches that will be discussed in subsequent chapters all deal with the determination of a backpack within a single frame. However for a video sequence we need more than one positive frame before we return an overall positive result. This section starts with a high level overview of the steps used in determining the classification of the whole video before a more detailed discussion in the next sub-section.

1. **Count Appearances of Backpack** – Each frame a backpack is present in increments a counter. This is only recorded if the location of the individual in this frame is within a certain

distance of their previous location, as a person can only move a relatively small distance between frames.

2. **Adjacent Frames** – In other situations, as indicated by the video above the backpack may be detected very well and give a continuous response, however it may only be visible to the detector for a relatively small amount of frames. In these cases once a consecutive number of frames had been detected with a backpack, the sequence was classified as a true positive. This minimum number of frames was given by the tuneable parameter MIN_FRAMES.
3. **Percentage of Appearance** – Some of the harder to detect backpacks within the test data would not trigger N consecutive frames. However there overall percentage of frames a backpack was detected in would still be quite high as indicated by sequence 769 in **fig. 20.** Hence as the video progressed, a high local average of detected frames was searched, for instance four out of five frames. The optimal percent was determined by the tuneable parameter FRAME_PERCENT.

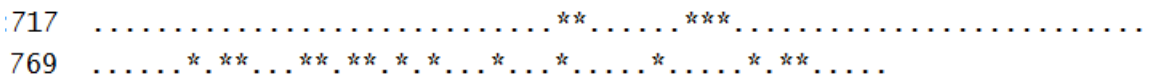


Figure 21: There two traces indicate all of the frames detected for videos 717 and 769 respectively. The dots indicate frames where a person has been detected but no backpack, the starts indicate frames a backpack was detected in.

4.5.1 GUI Window

Figure 22: Example of GUI window for approach five.

A GUI window was designed for approaches three and four and is shown above in *fig. 22*. The aim of this GUI was to demonstrate the system and the underlying detection methods rather than to provide a working interface for an end user. Several elements are common to both GUI designs:

- **A** – The input window shows a reduced resolution (640 x 360 pixels) version of the original camera input.
    - **Aa -** A red box indicating any detected people is overlaid on this image. If a strap is detected within this frame
    - **Ab –** If a strap is detected a green box is drawn around the strap
- **B –** A progress bar indicates the detection result for each frame.
    - **Ba** - Frames that have not yet been processed are shown in grey ().
    - **Bb** - If a person is not detected a frame is rendered as black after it has been shown.

    - **Bc** - If a person is detected but not a backpack the frame is rendered as red.
    - **Bd** – If a person is detected and a backpack the frame is rendered in green.
- **C –** The video number is displayed
- **D –** The frame length of the video is also displayed
- **E –** The annotated data for the whole video is also displayed as one of three values: No Backpack present, double strap backpack or single strap backpack.
- **F –** The number of frames a backpack has been detected in
- **G –** The first frame a backpack was detected in
- **H –** The last frame a backpack was detected in
- **I –** The range of frames a backpack was detected in
- **J –** The result of the system, if a backpack is detected it will be displayed in green

4.6 Summary

This chapter has documented the pre and post processing steps used by all of the systems. It is very important that the pre-processing steps are conducted as robustly and accurately as possible. Any error introduced at this stage will propagate throughout the system and reduce the ability of later steps to accurately determine if a backpack is present or not.

5 Approach One: Colour Space Clustering and Statistical Classification

This approach clusters the entire image in colour space and extracts connected regions to check if they are representative of the straps using statistical means.

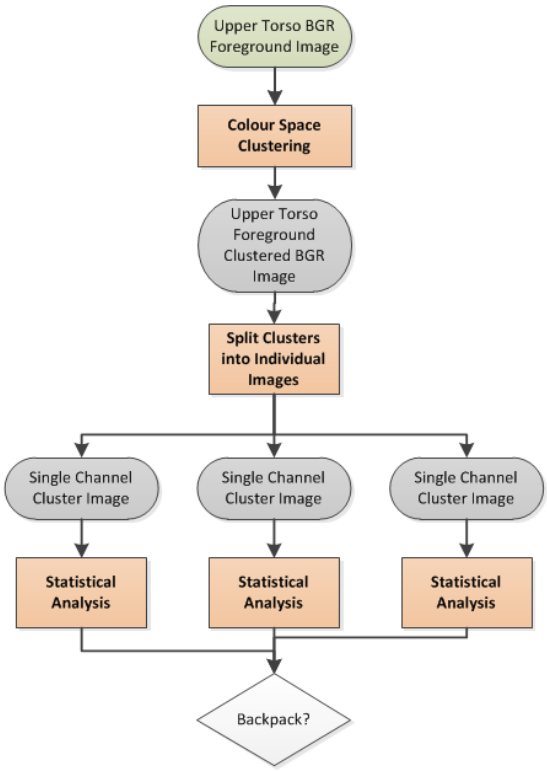


Figure 23: Flow Chart for Approach One

5.1 Design Overview

1. **Colour Space Analysis** – All of the pixels that lie within both the region of interest and the foreground mask are grouped into an appropriate cluster depending on their RGB values.
2. **Connected Region Analysis** – Each of the clusters is considered separately and connected pixels within each cluster are all grouped into regions.
3. **Region Statistical Analysis** – Every region within the image is statistically analysed to determine if it is a strap or not based on its length to width ratio and location within the image.

5.2 Colour Clustering

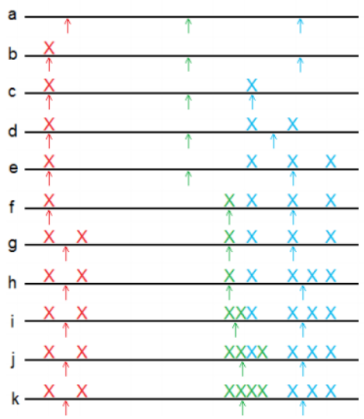


Figure 24: K-Means cluster centres changing over iterations, image from [1]

For approach one K-Means clustering was applied to the full resolution three channel colour image. K-Means clustering is a method for grouping data points with similar properties together [34]. It is often used in image processing to reduce the number of colours in an image as it is easier to retrieve information if it is in a more ordered format. For instance, take a three channel eight bit colour image, it has 256^3 possible colour combinations. This is simply too much information and we may wish to reduce the number of colour present. K-Means can be used to reduce the number of colours present to two, three, four or any other number we wish. K-Means works as follows:

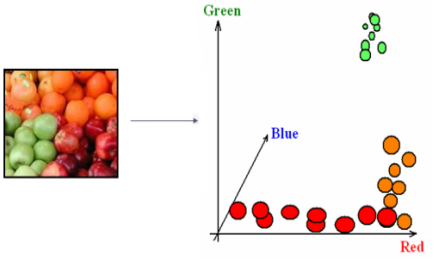


Figure 25: 3D Colour Based Clusters, [1]

1. The initial cluster centres are seeded using either random values or user supplied information.
2. Data points are grouped according to their closest cluster centre.

3. The cluster centres are then re-positioned to be at the centre of all the points assigned to them.
4. Step 3 is repeated until the centres stop moving by a pre-defined level of accuracy or a maximum number of iterations is exceeded.
5. The compactness of the clustering is determined as: $\sum_i \|samples_i - centers_{labels_i}\|^2$

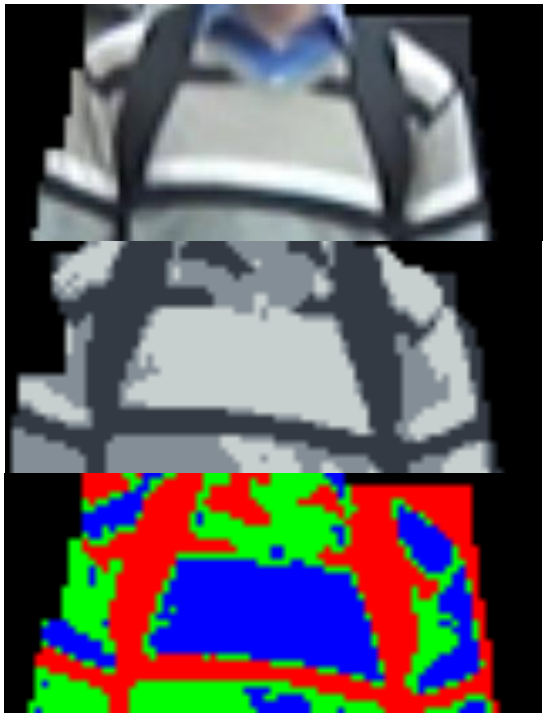


Figure 26: Original image on top clustered into three clusters. The middle image shows the clusters using representative colours while the bottom image uses more contrasting colours.

While the algorithm has been described above as working on three channels in RGB space it can also operate on any number of channels in several colours spaces. These colour spaces are discussed in *appendix A2*. Several factors had to be taken into account when tuning the K-Means algorithm:

- K-Means is an iterative process as the centres of the cluster move on each iteration, to prevent the process from continuing indefinitely two termination criteria are set. The first was an accuracy based criteria with the process terminating when the cluster centres stopped moving by more than a specified level of accuracy. Due to the low number of clusters this could be quite coarse and was set to 10.0. The second was an overall cap on the number of iterations. A value of 10 was chosen to prevent excessive clustering from slowing down execution time.
- A bad set of initial cluster centres can prevent the algorithm from returning the correct values. Hence random selection of initial clusters is not a good idea. Initial seeds were generated using the k-means++ method developed by [35] and explain in *appendix A2.*
- The algorithm can be run several times with different k-means++ seeds generated each time. The run with the lowest level of compactness would be used.
- The number of clusters could also be varied. From the process of applying K-Means to the overall image it was found that as backpack straps are usually of one or two uniform colours a low number of clusters are ideal. With more than five clusters the boundaries between straps and underlying garments tended to be given its own cluster which would trigger multiple responses for a strap. Even worse was that the boundaries between different garments could be given their own cluster which tended to trigger false positives.

5.3 Statistical Analysis

In Approach One each of the clusters was segmented into its own image. In each of these images the pixels were analysed to find connected region as indicated in *fig. 27*. Each of these connected regions could then be statistically analysed to see if it fitted the profile of a strap as follows:

1. A minimum bounding rectangle was applied to all connected regions. From this the height width ratio was calculated and compared against a threshold.
2. If the centre point of the connected region is left of the centre line and it passed step 1 it is placed into a bin of left straps, alternatively if it's centre is right of the centre line it is placed in a bin of right straps
3. All combinations of left and right straps are compared to see if they are positioned symmetrical between both sides of the image and in a sensible location corresponding to likely strap locations.

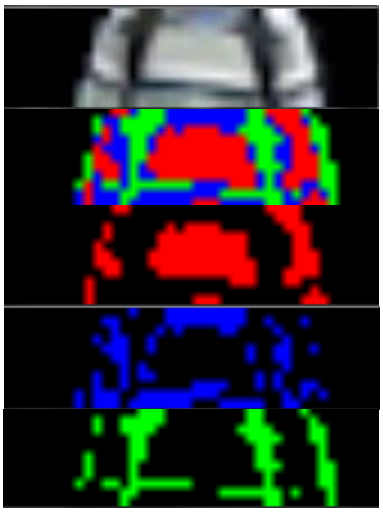


Figure 27: Clustered Region split into separate channels for each cluster.

5.4 Results

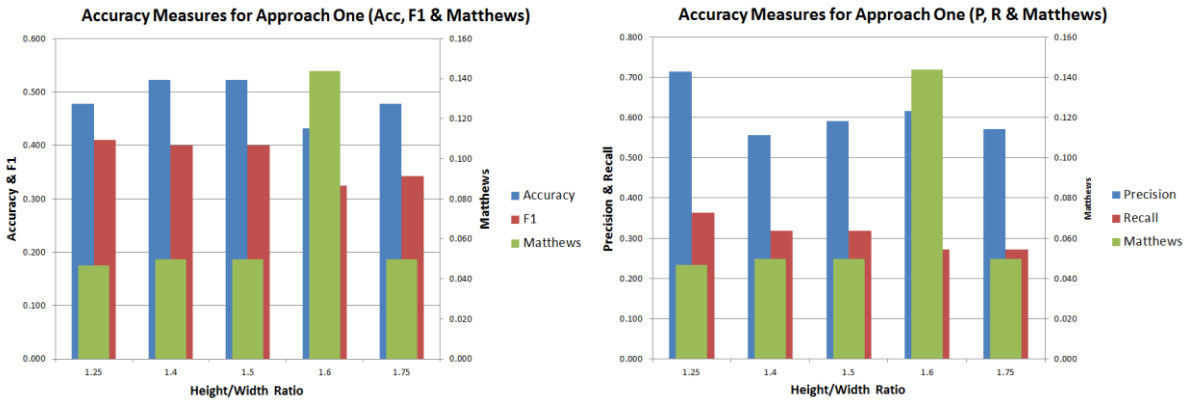
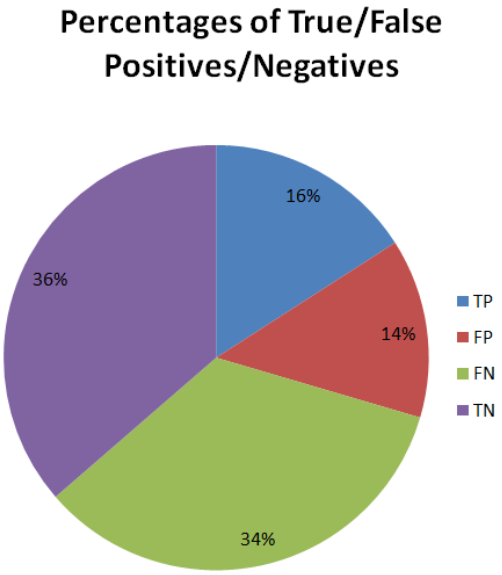


Figure 28: The height/width ratio is varied, a ratio of 1.4 is found to be the optimal. (Please note the Matthews is on the right y axis).



TP	FP	FN	TN	Recall	Precision	Accuracy	F1	Matthews
7	6	15	16	0.304	0.538	0.523	0.389	0.050

Figure 29: Best Results for Approach One

Approach One has the lowest accuracy, F1 and Matthews scores of any of the approaches. As can be seen in *fig. 28* it has low detection rate (TP & FP). This is reflected in *fig. 29* which shows that the recall value is very low at 34%. The precision is also low as barely over half of the positives are correct. There were two tuneable parameters used in approach one:

- **CONNECTED_REGIONS** – This governs the threshold for which the height/width ratio is compared against. A value of 1.4 was found to be the optimal.
- **CLUSTER_COUNT** – This governs the number of clusters used in the K-Means process

5.5 Evaluation

Figure 30: Too few clusters (left). Too many cluster (centre). Straps extracted with additional regions (right).

Given such low scores it is clear this approach is a failure, the results indicate it is only slightly more accurate than simple random classification. The main reason for this failure is that clustering has been applied over the whole image. Depending upon the diversity of the colours within the image this can result in there being too few clusters causing the straps to be clustered with neighbouring parts of the garment. Alternatively there may be too many clusters leading to the top and bottom of the straps being placed in different clusters. The clustering was heavily influenced by illumination changes as well as colour changes. This factor caused incorrect segmentation of the straps from the image as indicated in *fig. 30.* Hence the height width ratio would fail to evaluate correctly. As has been documented in *appendix A2* attempts were made to classify images based on different colour spaces however these attempts proved un-successful.

5.6 Summary

This approach was the least promising of any approach and in the authors opinion is not suitable for use alone within a system. Some of the statistical means employed may have a place supporting other methods but not as a stand along system. To try and reduce the systems susceptibility to illumination changes clustering should in future be applied to a more localised area.

6 Approach Two: Edge Gradient and Orientation Analysis

This approach is based on edge analysis of the image and looks for orientation combinations that signify a strap being present.

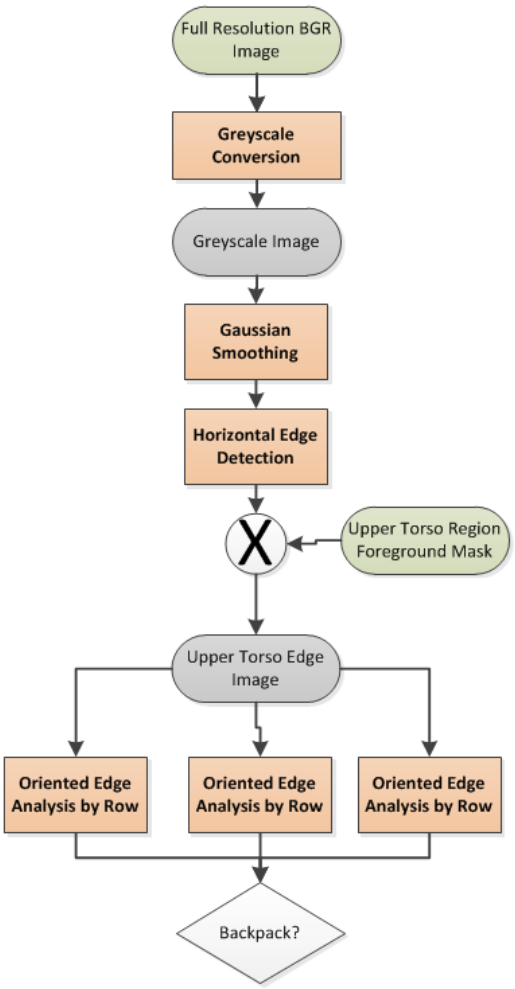


Figure 31: Flow chart of Approach Two

6.1 Design Overview

1. **Greyscale Conversion** – The image is converted into greyscale, the edge detection used does not take account of colour information.
2. **Gaussian Smoothing** – Edge detection works better when the noise levels within the image are reduced, hence Gaussian Smoothing is applied to the greyscale image.
3. **Horizontal Edge Detection** – Backpack straps are usually represented by vertical lines hence we are only interested in detecting edges in the horizontal direction. Edge detection is applied to the whole image as the gradient values for pixels at the edges of images are calculated using pixels from the surrounding region, parts of which will be outside the mask.
4. **Non-Maxima Suppression** – NMS is applied to the image in the horizontal direction only to give a single pixel wide edge which can be used to pin-point its location.
5. **Thresholding** - The edge image is then thresholded to remove weak edges below a certain magnitude value. This avoids secondary edges from disrupting with detection.
6. **Foreground Mask** – The foreground mask and region of interest masks are then applied to the edge image so that we are only taking into consideration edges within the upper torso region.
7. **Row Analysis** – For rows at evenly spaced intervals along the image we count along the row looking for edges and their orientation. If the correct combinations of negative and positive orientations are encountered that indicate two straps and arm edges we conclude that there is a backpack on that row.
8. **Whole Image Analysis** - If a sufficient number of the rows agree we conclude that there is a backpack present in this frame.

6.2 Edge Detection

The purpose of this step is to take a greyscale version of the full resolution input image and return a map of the main edges for the entire image. There are numerous full edge detectors already implemented in OpenCV however they all detect gradients based on both the X and Y direction. For this application the assumption is being made that the input image will always be oriented so that it is the correct way up. Hence when searching for backpack straps we are only interested in edges oriented along the y-axis with a gradient change in the x-direction. Hence a new edge detector was implemented loosely based upon the Canny edge detector in the x-direction only.

Figure 32: Top Row: Colour Upper Torso Region, Greyscale Upper Torso Region, Gaussian Bluring Applied Middle Row: 1st Derivative Sobel in x-direction, Thresholding applied, Non-Maxima Suppression Applied

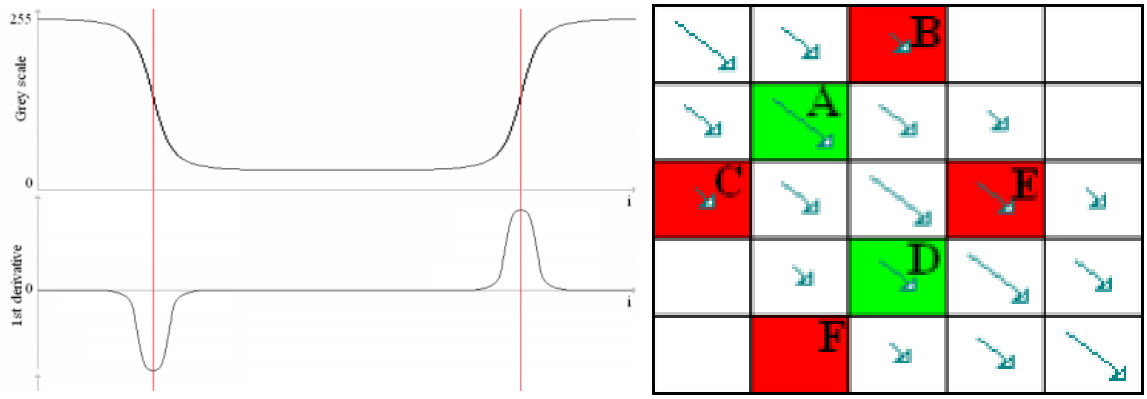


Figure 33: Left: Top row shows intensity of pixel and bottom row shows gradient intensity. Right: Green Pixels have strongest response and are kept, while red pixels are suppressed during NMS. Both images from [1]

6.2.1 Gradient Analysis

This sub-step takes the input greyscale image and returns the gradient in the x-direction across each pixel. Before this takes place Gaussian Blurring is applied as indicated in the top right of *fig. 32* to the greyscale image. This reduces noise due to wrinkles in the clothing, small logos, buttons and

other items as can be seen for the logo in the top right of the torso. Eradicating these sources of short edges enables us to concentrate our analysis on only the more prominent edges.

A first derivative edge detector was chosen as it gave a high enough level of accuracy for the edge while also giving directional information about the pixel. A second derivative detector in theory can give a more accurate location but does not contain directional information about the edge. This is computed using a single Sobel mask as indicated in *fig. 32*. For each pixel the mask is centred on that location and all the weights applied to the surrounding greyscale pixels and summed up to obtain the gradient intensity of that pixel. By experimentation, I found that the best size for the sobel mask was 7x7 as indicated in *fig 34*.

A signed Sobel was used, which gave us both the directional and magnitude responses for the gradient within the image, as can be seen in the bottom left of *fig. 32*. The direction information was preserved as the two edges of a strap tend to be in the opposite direction. A threshold was then applied to this gradient image which eliminated all pixels below a certain magnitude. The result of this can be seen at the centre of the bottom row where only strong responses remain. In effect we have now isolated the edges.

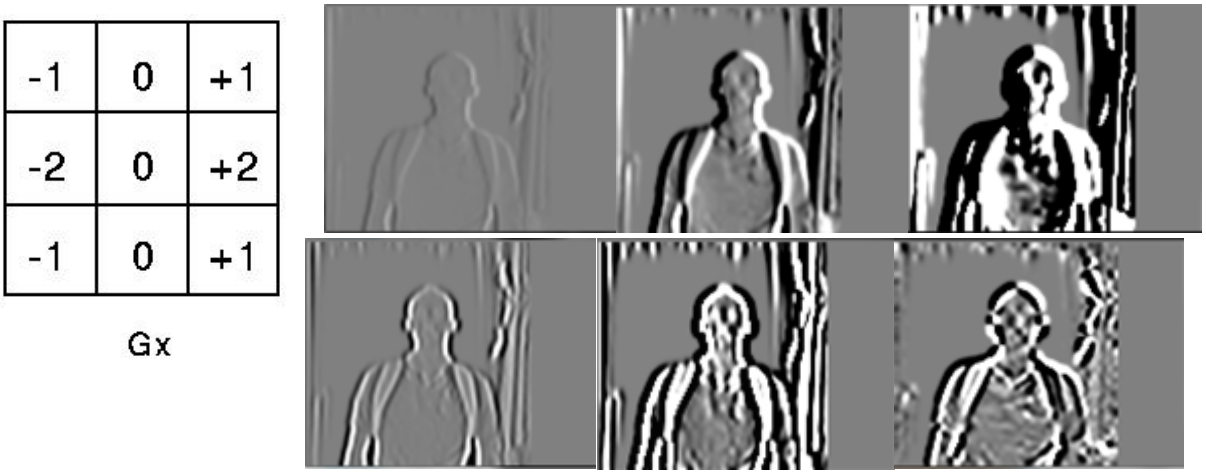


Figure 34: Left is a Sobel window for the x-direction, image from [1]. Right are six gradient images creating using a 1st derivative Sobel (top row) and 2nd derivative Sobel (bottom row). Both are in the x-direction only using window sizes of 5, 7 and 9.

6.2.2 Non-Maxima Suppression

Non-Maxima suppression was then employed in the x-direction only. As we were only using one edge this could be simplified compared to the conventional NMS algorithm. We could look at all pixels connected horizontally in a line and select only the strongest response. This is shown in *fig. 34* for a 45 degree orientation, this would take place in the x-direction on in this solution. This reduced the edges to only one pixel wide which gave us a local maxima where we could define the edge location as can be seen in the bottom right of *fig. 32*.

6.3 Orientation Analysis

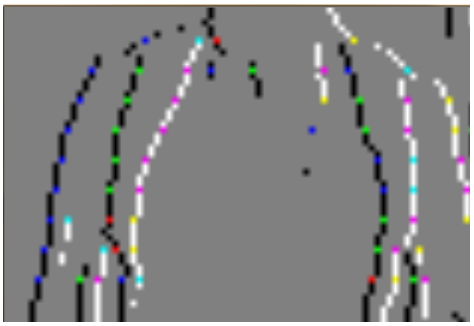


Figure 35: Coloured pixels represent edges encountered as rows are counted across

The classification system employed in approach two is quite simple, counting across an arbitrary number of rows in the oriented edge image. On each of these rows it is looking for 3 edges with a specific combination of orientations. These edges represent the edge of the arm and the two edges of a strap as all three of these give a strong response in most images. The possible orientation combinations that satisfied these conditions are show in *fig. 35*. The method required the finding of two straps on both sides of the image before giving a positive classification for that row. 10 rows were checked per image and the number of positive rows required for a positive classification could be varied between 6 and 9.

6.4 Results

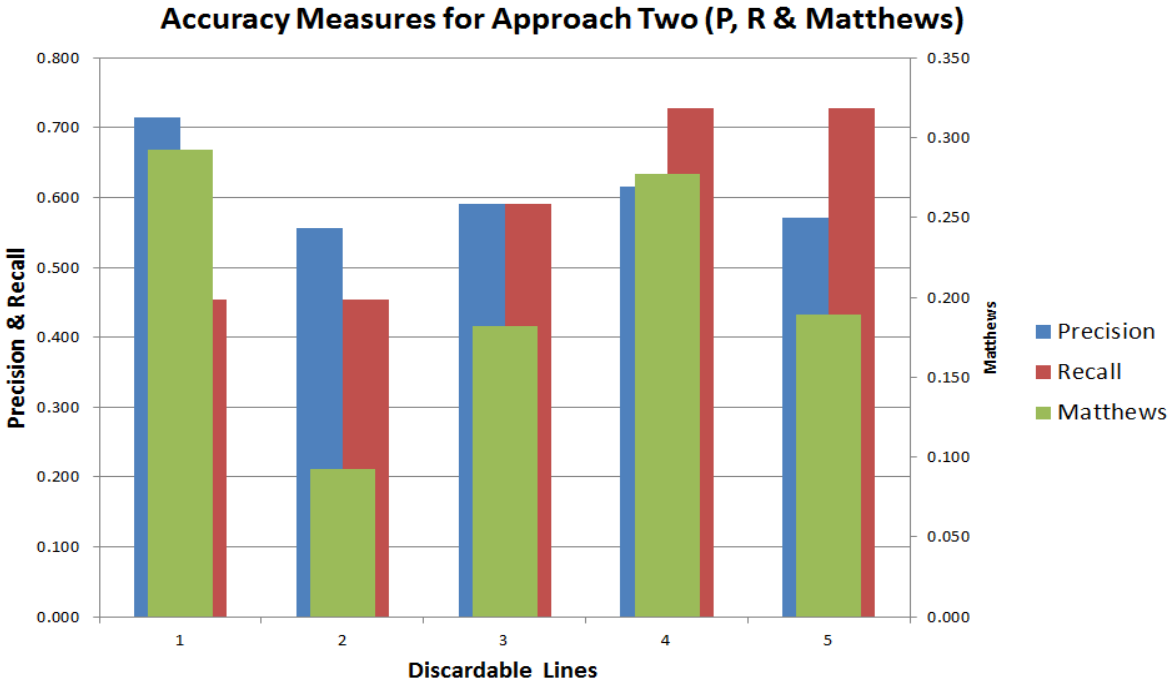


Figure 36

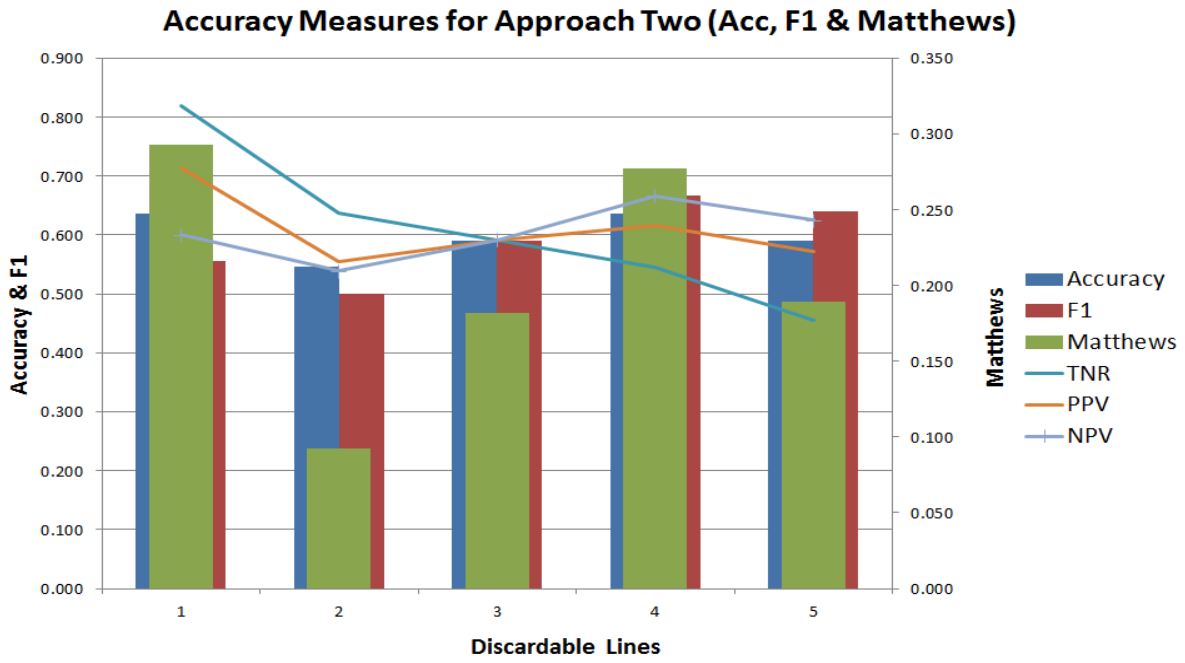
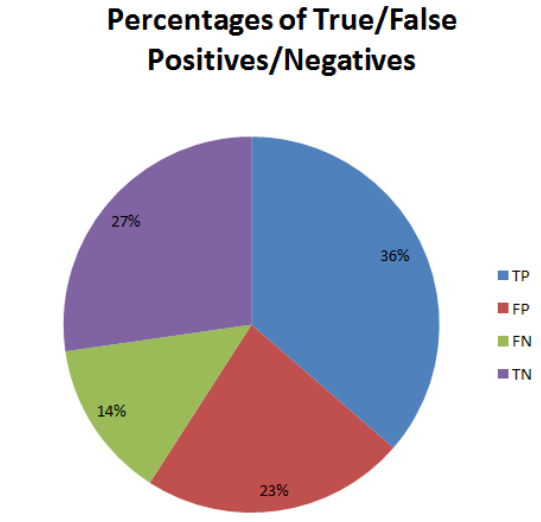


Figure 37: the number on the x-axis represents the number of negative rows that can be discarded while still giving a positive response for the detection of a backpack within the overall image. (Please note the Matthews is on the right y axis).



TP	FP	FN	TN	Recall	Precision	Accuracy	F1	Matthews
16	10	6	12	0.727	0.615	0.636	0.667	0.277

Figure 38: Best results for Approach Two

Tuneable Parameter:

- **LINE_COUNT** – This represents the number of lines that can be discarded from consideration.

Approach Two has a very high detection rate as indicated by the high recall value present in *fig. 36*. However precision is still low as this method is also detecting a lot of false positives. The accuracy scores are low but giving better results than approach one. Increasing the number of lines that may be discarded from the result has the effect of increasing the values of recall and precision as can be seen in the graph in *fig. 37*. The increase in recall makes sense as having more discards results in greater levels of detection. The increase in precision suggests that the ratio of correct detections to false detections is also increasing.

However the Matthews value starts to vary a lot, decreasing and then increasing. Accuracy and the F1 ratio tend to follow the trend of the Matthews value. (Please note that the Matthews is on the right vertical axis while all other values are on the left vertical axis.) To better understand these results and the varying Matthews value we need to look at the True Negative Rate (TNR), Positive Prediction Value (PPV) and Negative Prediction Value (NPV) as can be seen in the left graph. These can all be seen to decrease as the number of discards is increased. Particularly alarming is the rapid descent of the TNR. This suggests that increasing the number of discards is allowing significantly more false positives. It should be noted that the Matthews is much better than the other values at

taking into account the ratio between true and false positives as well as true and false negatives than any of the other values, so it's results should carry more weight.

6.4.1 Successes

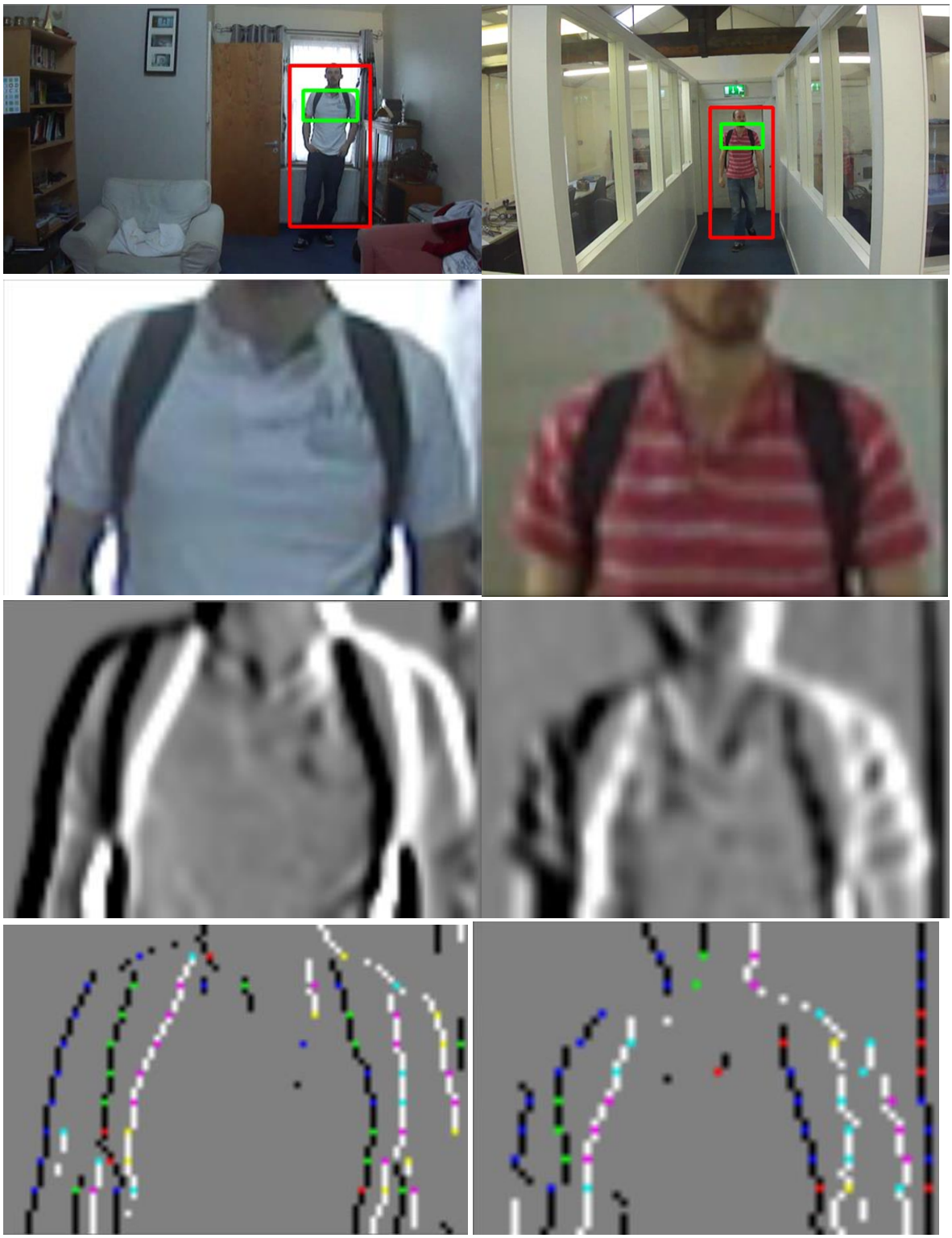


Figure 39: Successful Detection of backpacks using Approach Two

6.4.2 Failures

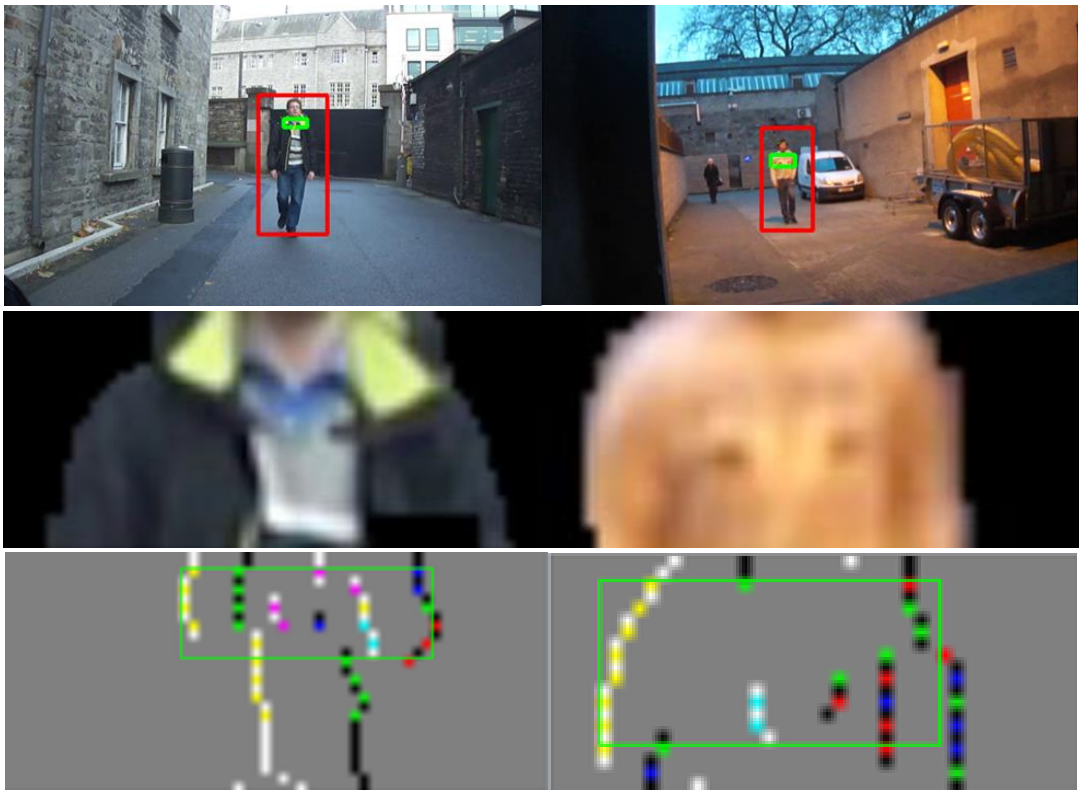


Figure 40: Failures of Approach two. Left show the lapels being detected as a backpack giving a false positive. Right shows shadows being detected as a False Positive

6.5 Evaluation

Generally this method's weakness is its reliance on the edges generated by the straps of the backpack and edges of the arm extending down the full length of the upper torso region. This is often not the case and false negatives are caused when several of the rows do not detect a backpack. Increasing the number of discards to compensate for this, weakens the system, as it will now rely on relatively few rows for classification. In addition there is no check to ensure continuity between what is being detected in one row or another. Hence as can be seen in the right sequence of *fig 40* some of the short edges introduced by the blurred logo in the upper torso region are being detected as parts of straps. In addition, in the left sequences, parts of the yellow jacket lapels are also trigger a response.

6.6 Summary

This method presents results that are better than approach one's results but still not great. This method has a fairly high detection rate and therefore has a problem with detecting backpacks when they do not exist. Potential improvements to the system would be to try and increase the continuity between what is detected on each row. Currently the system only looks for combinations of edges but does not account for their orientation relative to each other or if the same edges are being detected in every row.

7 Approach Three: Parallel Edge Analysis

The aim of this approach was to try and capture more information about pairs of parallel edges that form potential candidates for straps. Better classification could then be made using the full length of the edges as well as their orientation being taken into account. This is in contrast to approach two which only takes localised snapshots on pre-defined rows. As the initial stages of approach two and three are identical the reader should read section 6.1 from approach two's chapter before proceeding further.

The chapter will start off with a high level overview in section 7.1. The method of finding parallel contours will be discussed in section 7.2.

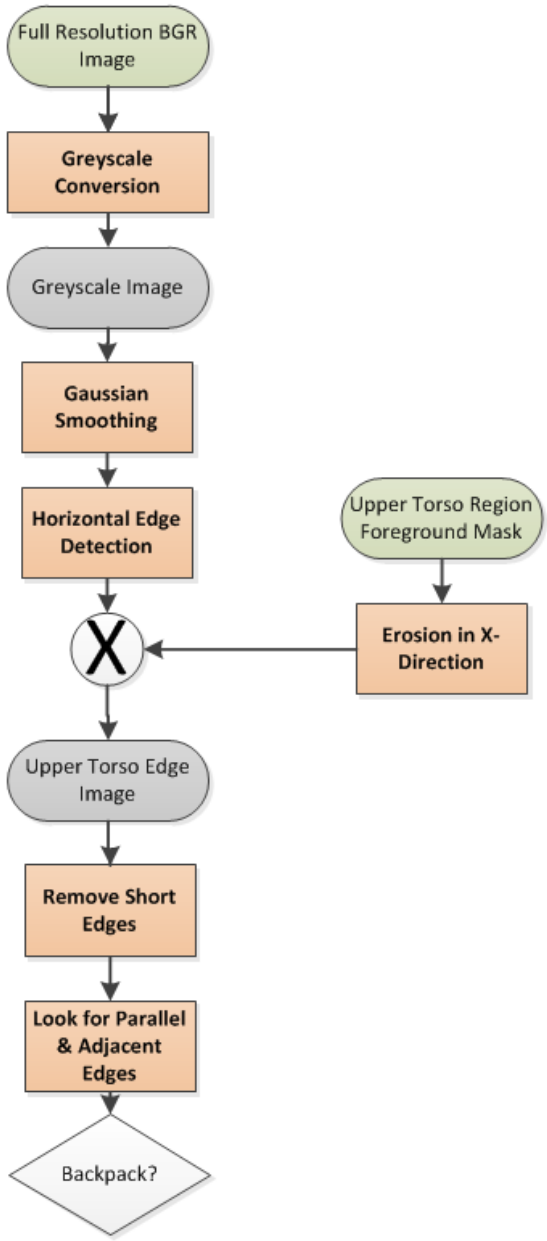


Figure 41: Flow Chart for Approach Three

7.1 Design Overview

Steps 1.5 are identical to those found in approach two.

1. **Greyscale Conversion** – The full resolution input image is converted into greyscale.
2. **Gaussian Smoothing** – Edge detection works better when the noise levels within the image are reduced hence Gaussian Smoothing is applied to the whole input image.
3. **Horizontal Edge Detection** – Backpack straps are represented by vertical edges hence we are only interested in gradient changes in the horizontal direction. Hence edge detection will be along the x-axis only. This step is applied to the whole image as the gradient values for pixels are calculated using neighbouring pixels that may be outside of the upper torso region.
4. **Non-Maxima Suppression** – NMS is applied to the image in the horizontal direction only

to give a single pixel wide edge which can be used to pin-point the location of the edge.

5. **Thresholding** - The edge image is then thresholded to remove weak edges below a certain gradient value. This avoids secondary edges from disrupting calculations.
6. **Foreground Mask** – An eroded version of the foreground mask combined with the region of interest mask was applied to the edge image. The erosion prevented edges at the edge of the region of interest from being considered as straps.
7. **Connected edges** – This method searches for long chains of connected edge pixels. It extracts their length, orientation and location of both ends.
8. **Remove Short Edges** – Edges below a certain length are not likely to be considered backpack straps and are removed from consideration.
9. **Parallel Edges** – If two edges have their ends within a specified distance of each other and have an orientation that is within a certain angle of each other they can be considered as parallel. These parallel edges are designated as a strap.
10. **Backpack Detection** – If we have two parallel straps in roughly symmetrical locations on either side of the upper torso region we consider this to be a backpack.

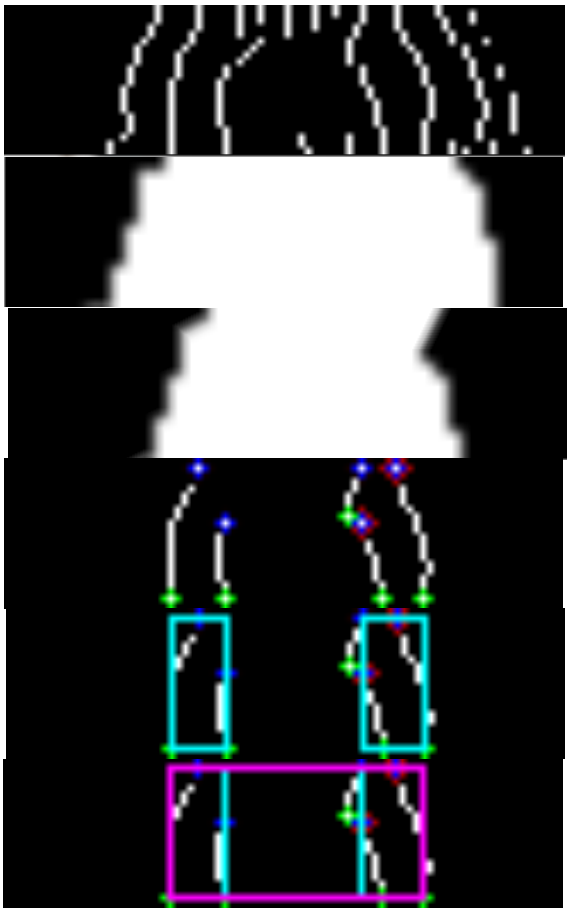


Figure 42: From Top to Bottom: Original Upper Torso Region, Greyscale, Gradient Image, Edge Image, Upper Torso Region, UTR eroded in the x direction, edges filtered for sensible strap locations, parallel pairs identified, symmetrical parallel pairs identified

7.2 Parallel Contour Extraction

This step takes the x-direction edge image visible in *fig. 42* and applies the upper torso region foreground mask. This mask is shrunken slightly in the x-direction to remove any edges that belong to the outside of the arm. This method also disregards the orientation information about edges. This was found to be un-reliable in approach two and for a mere two edges will not provide much useful information.

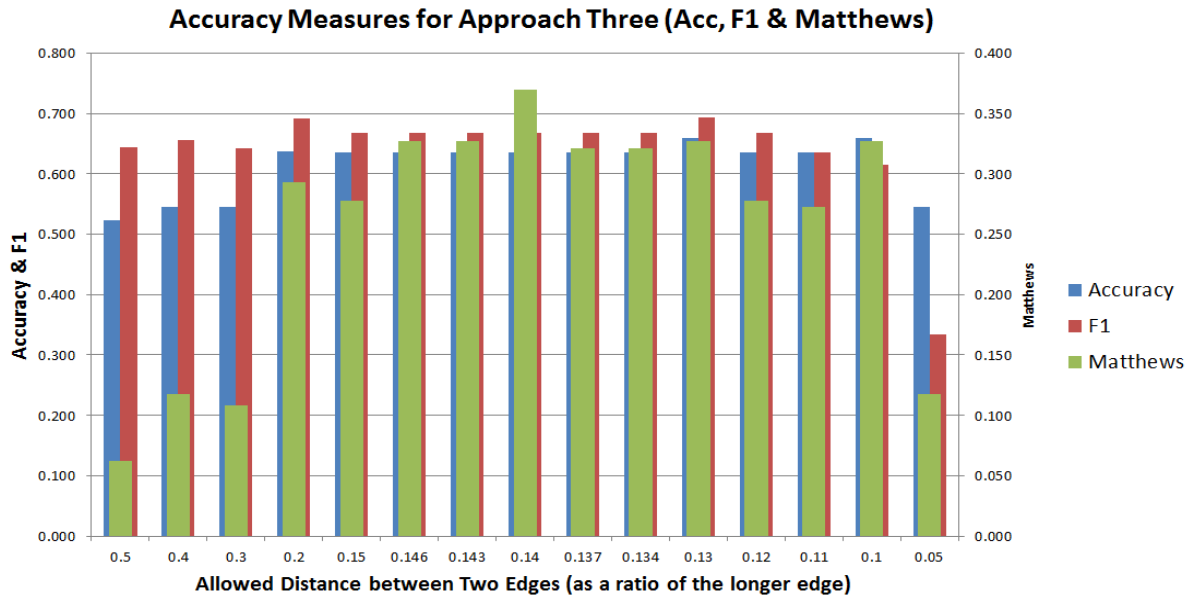
Each chain of pixels that is connected together is considered as a contour representing an edge. Extra short edges are often produced by lapels and logos**.** These short edges are eliminated from consideration, by simply disregarding any contour with a length less than a certain percentage of the overall height of the upper torso region of interest.

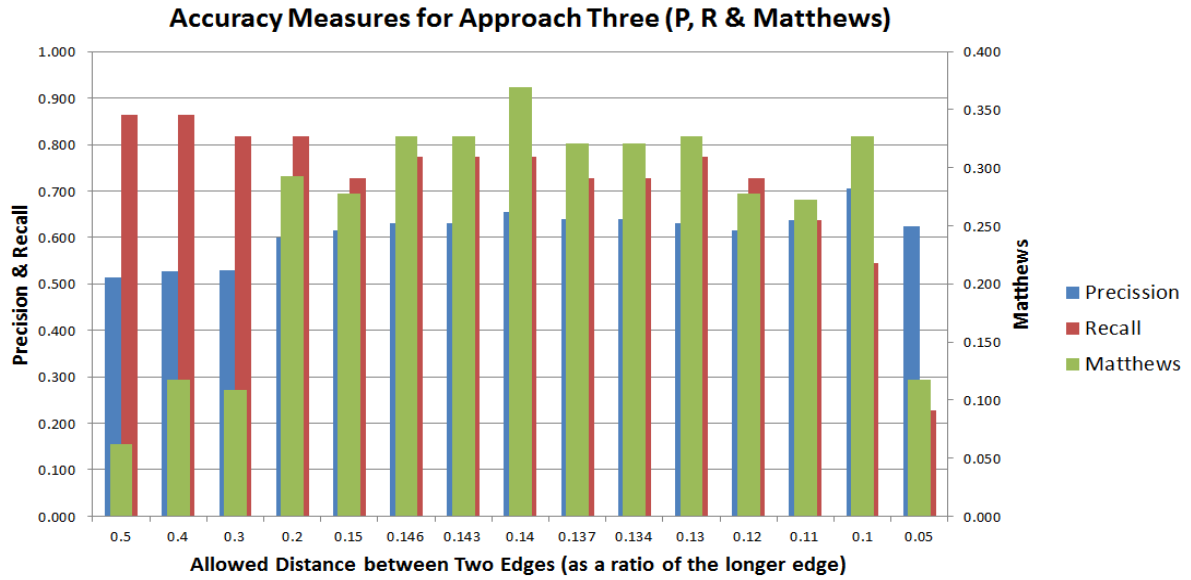
7.3 Symmetry Analysis

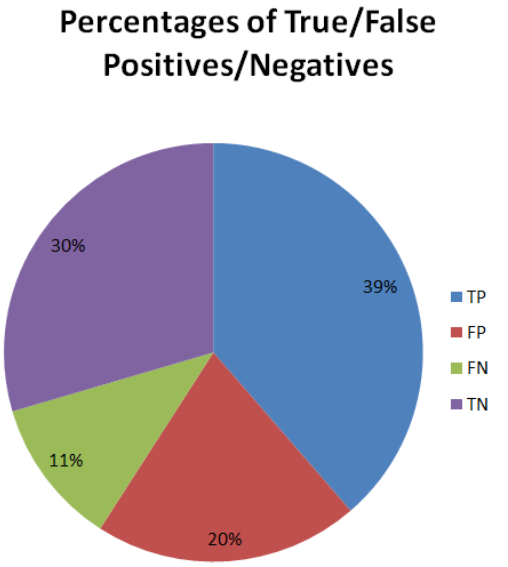
All of the contours are compared to each other looking for pairs with a similar orientation or the inverse orientation (+/- 180 degrees). The orientation is determined as the angle of a line between the start and end point of the contour relative to the x-axis. All pairs that have a similar orientation are checked to see how close together they are relative to their total length. If they are close enough the pairs are considered as a strap object and stored for further analysis.

Approach three finishes off by searching for two symmetrical straps on either side of the upper torso region. If two are found of roughly similar size they are considered as a found backpack.

7.4 Results







TP	FP	FN	TN	Recall	Precision	Accuracy	F1	Matthews
17	9	5	13	0.773	0.654	0.682	0.708	0.370

Figure 43: Best results for Approach Three

Approach Three contained several tuneable parameters:

- **DIST** – This governed the allowable distance between two contours. If this value was exceeded the contours where no longer considered as a potential strap. The optimal value for this was found to be 0.14 times the length of the longer contour.
- **ORI** – This governed the allowed orientation distance between two contours. If this value was exceeded the contours were no longer considered as a potential strap.
- **SOBEL_MIN_RANGE** – This governed the sensitivity of the thresholding of the gradient image. It was found that for this solution a relatively high value was desirable for the threshold to enable consideration of only the best candidates for edges.
- **X_ERODE** – This governed the amount by which the upper torso region was eroded in the x-direction. Once again a surprisingly large value was favoured with the ideal erosion being around 0.1 times the width of the upper torso region on either side.

7.5 Evaluation

Taking into account the full length of the edge as well as its orientation has greatly increased the accuracy of the system. It ensures that when the system indicates a strap it has at the very least located two parallel edges. This alleviates previous failures caused in approach two by random edges causing an arbitrary number of rows to appear to have straps on them.

7.6 Summary

Approach Three's strength was its high detection rate and low false negative rate. The main weakness was the high level of false positives caused by the high detection rate. This requires setting thresholds to high values to try and reduce the level of false positives. This has the negative factor of introducing more false negatives as backpacks are missed. However even when these two issues are taken into consideration, this approach has still had the best results yet and is considerably better than simple random classification.

To improve this system will require incorporating an additional method to try and reduce the high rate of false positives enabling the lowering of the Sobel edge detection threshold.

8 Approach Four: Parallel Edge and Colour Space Analysis

This is an extension to approach three, to try and resolve the weaknesses of that approach. Approach Three had a tendency to miss certain backpacks as they would not satisfy one or other of the requirements. However if these requirements were lowered it resulted in too many false positives. To enable the lowering of these requirements without compromising on the false positive rate the colour histograms of straps and surrounding regions are analysed for similarity. It is assumed the reader is familiar with approach three before they commence reading.

This chapter starts off with a high level overview of the design. The theory of the colour histogram analysis is discussed in section 8.1. Results are presented in section 8.2, evaluated in 8.3 and conclusions drawn in 8.4.

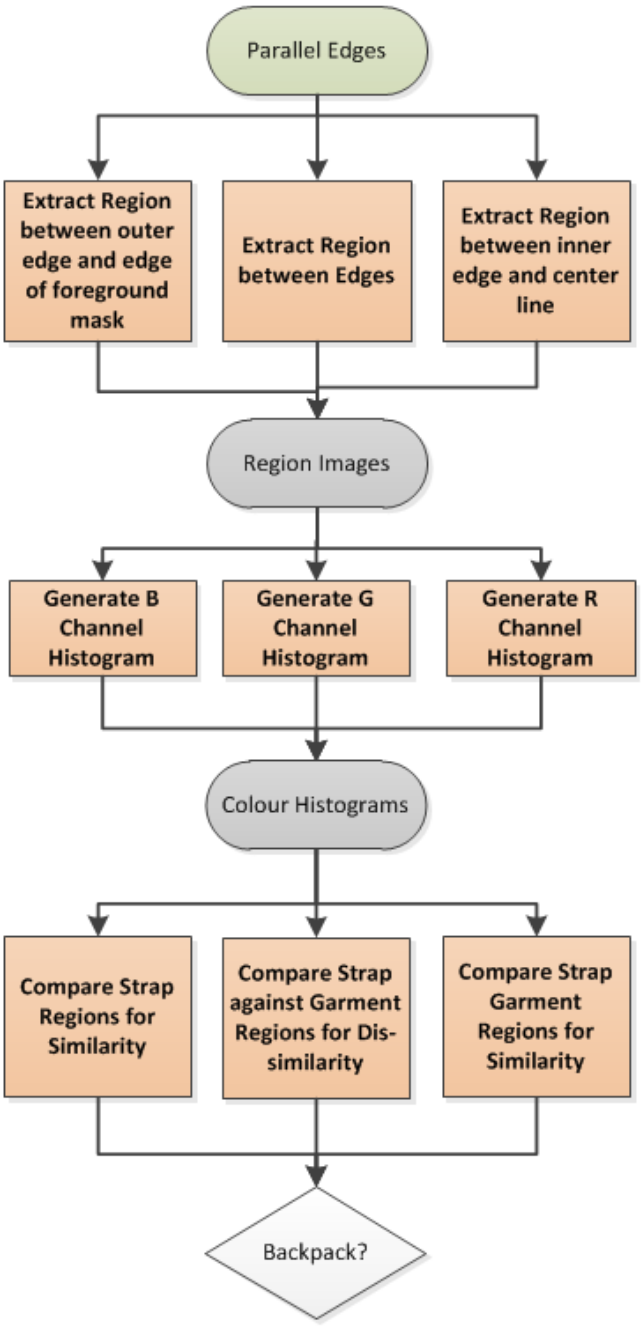


Figure 44: Flow chart for Approach Four

8.1 Design Overview

1. **Join Parallel Straps** – Contours were drawn around parallel strap pairs to define the strap region.
2. **Designate Garment Region** – The region between the inner edge of the strap and the centre line of the upper torso region is defined as the garment region. This was combined with its neighbour for the opposite strap to give one garment region and two strap regions.
3. **Colour Histogram** – A colour histogram of each region was generated.
4. **Comparison of Histograms** – These histograms were then analysed to see how similar they were. Ideally strap regions would be similar to each other as would non-strap regions to each other. However strap and non-strap regions should present a very low level of similarity. If these values were above and below certain thresholds respectfully the system would confirm the presence of a backpack.

8.2 Colour Histogram Analysis

The reasons behind using colour histograms are simple. Edges can be created by shadows, zips and other artefacts that could trick the system into thinking a strap is present. However a real strap is likely to have a different colour composition to the underlying garment. Hence comparing colour histograms should eliminate false potential straps.

When defining the strap region the two contours representing either side of the strap are joined together. Inevitably some of the edge pixels would be from the garment due to the detected edge locations not being perfect. Hence the strap regions were slightly eroded to eliminate these discontinuities.

The centre regions are defined as all pixels between the inside edge of strap and the centre line for all rows that have a strap. When creating colour histograms the values for all pixels channels. For each of them a histogram is created within each region. Three colour histograms are generated, one for the left strap, one for the right strap and one for the centre region.

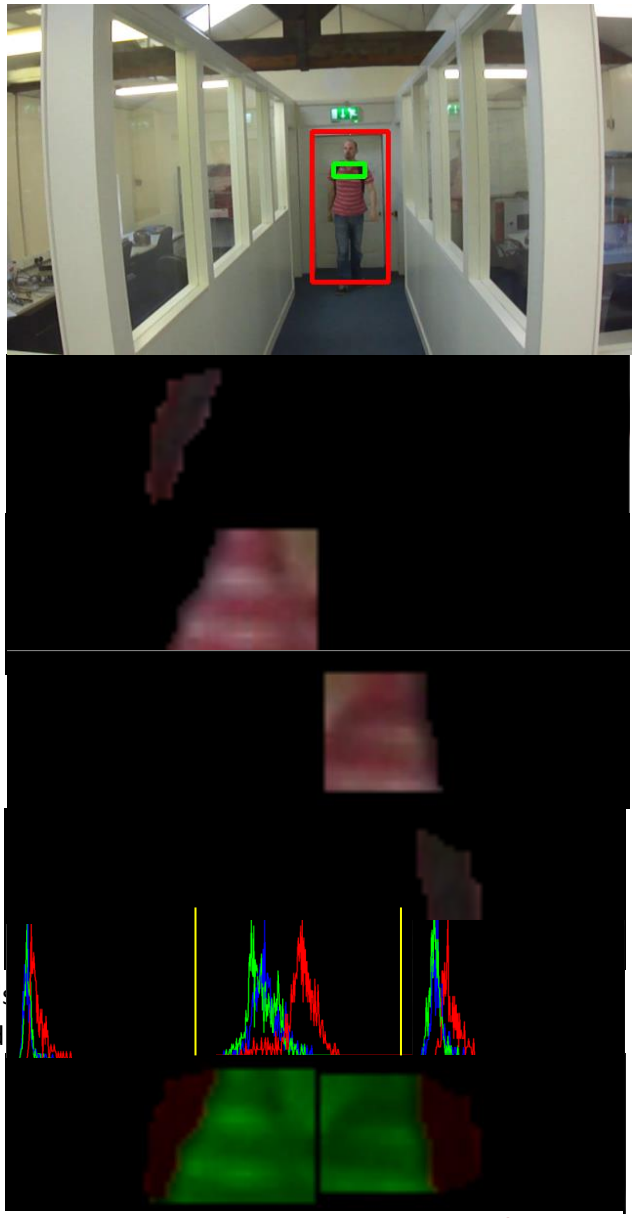


Figure 45 From Top: Upper Torso Region, potential straps in cyan and potential backpack in magenta, left strap, left garment, right garment, right strap, green indicates garment region and red indicates strap region, colour histograms for left strap, centre garment and right strap region.

8.3 Comparison

Unlike approach three this solution does not look for symmetrical pairs of straps. All potential straps on the left side of the image are compared to all potential straps on the right side of the image.

The colour histograms are compared separately for the red, green and blue colour channels. The comparison is made using a simple correlation calculation between the two histograms:

$$Corr(H_1, H_2) = \frac{Cov(H_1, H_2)}{\sigma_{H_1}\sigma_{H_2}} = \frac{\sum_I (H_1(I) - \mu_{H_1})(H_2(I) - \mu_{H_2})}{\sqrt{\sum_I (H_1(I) - \mu_{H_1})^2 \sum_I (H_2(I) - \mu_{H_2})^2}}$$

Alternative comparisons were also tried such as the Chi-Square, Intersection and Bhattacharyya distance. However the correlation co-efficient was found to give the best classification value after experimentation. If the co-efficient returned by the comparison between the left and right strap regions is greater than both the co-efficient between, a) left strap and central garment region as well as b) right strap and central garment region, the presence of a strap is confirmed. If multiple potential backpacks are located within the upper torso region, the one with the greatest difference between the co-efficients will be selected.

8.4 Results

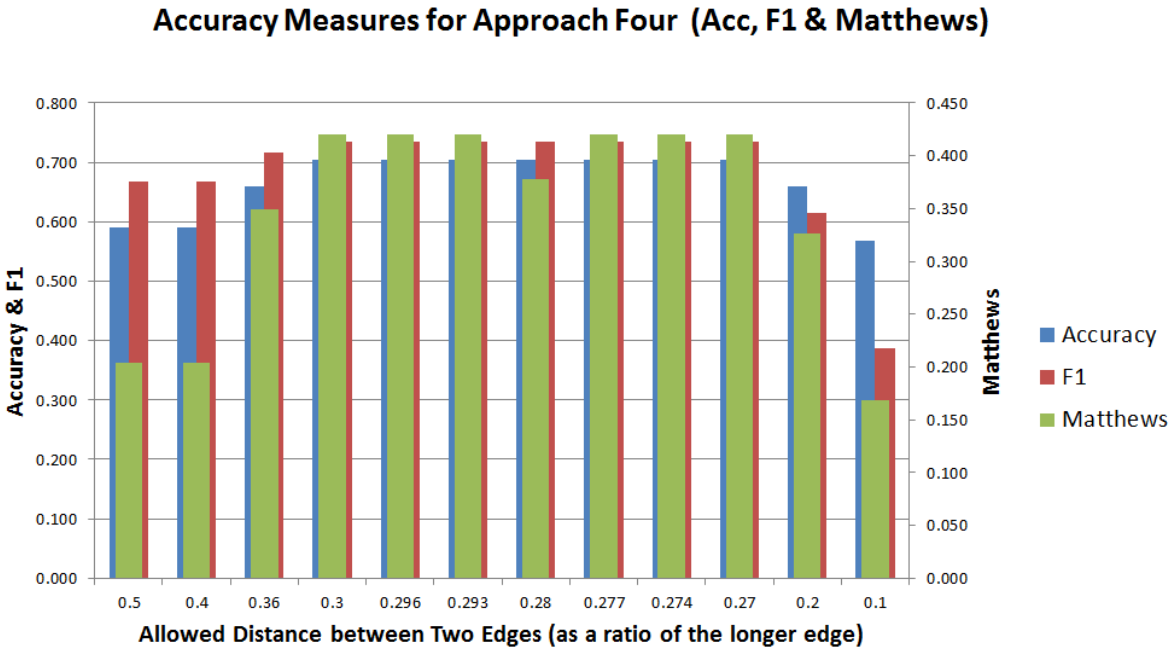


Figure 46

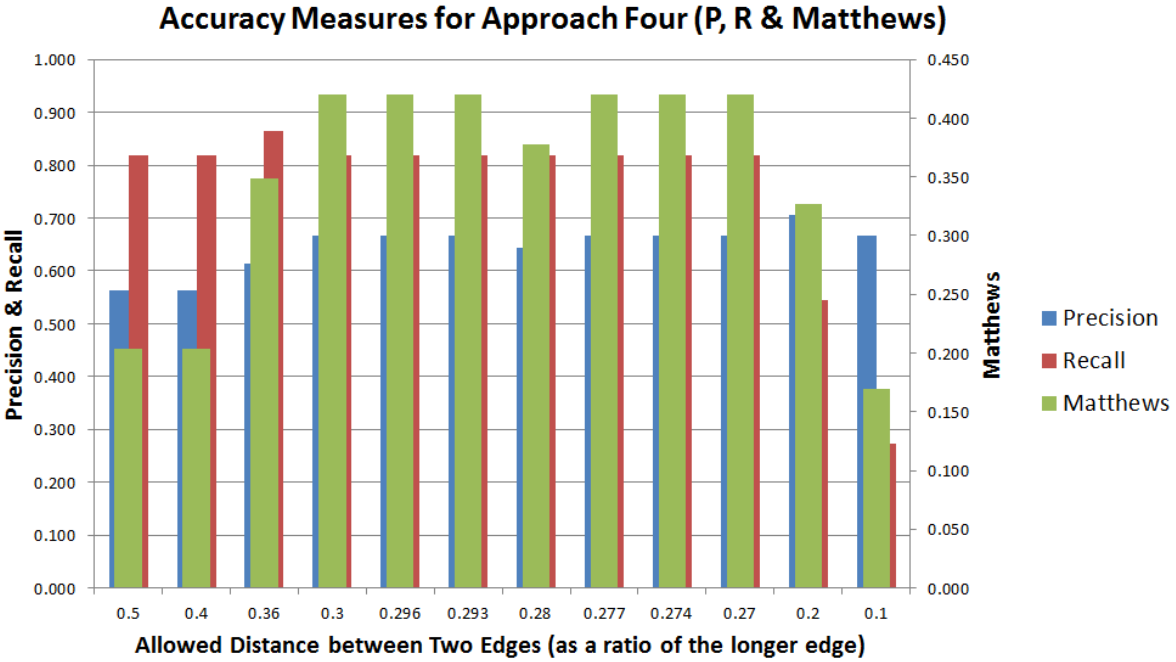
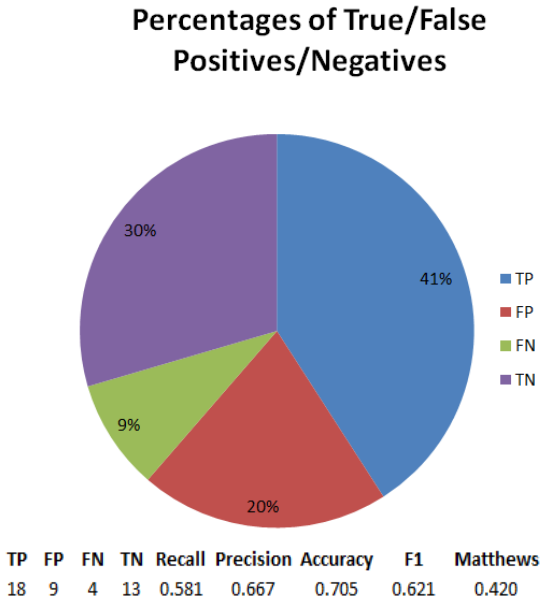


Figure 47



TP	FP	FN	TN	Recall	Precision	Accuracy	F1	Matthews
18	9	4	13	0.581	0.667	0.705	0.621	0.420

Figure 48: Best Results for Approach Four

Approach Four retained the tuneable parameters used in Approach Three:

- **DIST** – This governed the allowable distance between two contours. If this value was exceeded the contours where no longer considered as a potential strap. The optimal value for this was found to lie between 0.27 and 0.3 for this approach, as opposed to 0.14 for approach three.
- **ORI** – This governed the allowed orientation distance between two contours. If this value was exceeded the contours were no longer considered as a potential strap. The optimal value for this was found to be the same as in approach three.
- **SOBEL_MIN_RANGE** – This governed the sensitivity of the thresholding of the gradient image. The optimal value of this was found to be significantly lower than in approach three.
- **X_ERODE** – This governed the amount by which the upper torso region was eroded in the x-direction. The optimal value was found to be very similar to the value used for approach three.

Approach Four can be configured to have high recall or precision values, but struggles to achieve both in the same solution. This is evident in **FIG. 45 where** the recall starts of high but the precision low as the minimum distance between parallel contours is small.

8.5 Evaluation

8.5.1 Successes

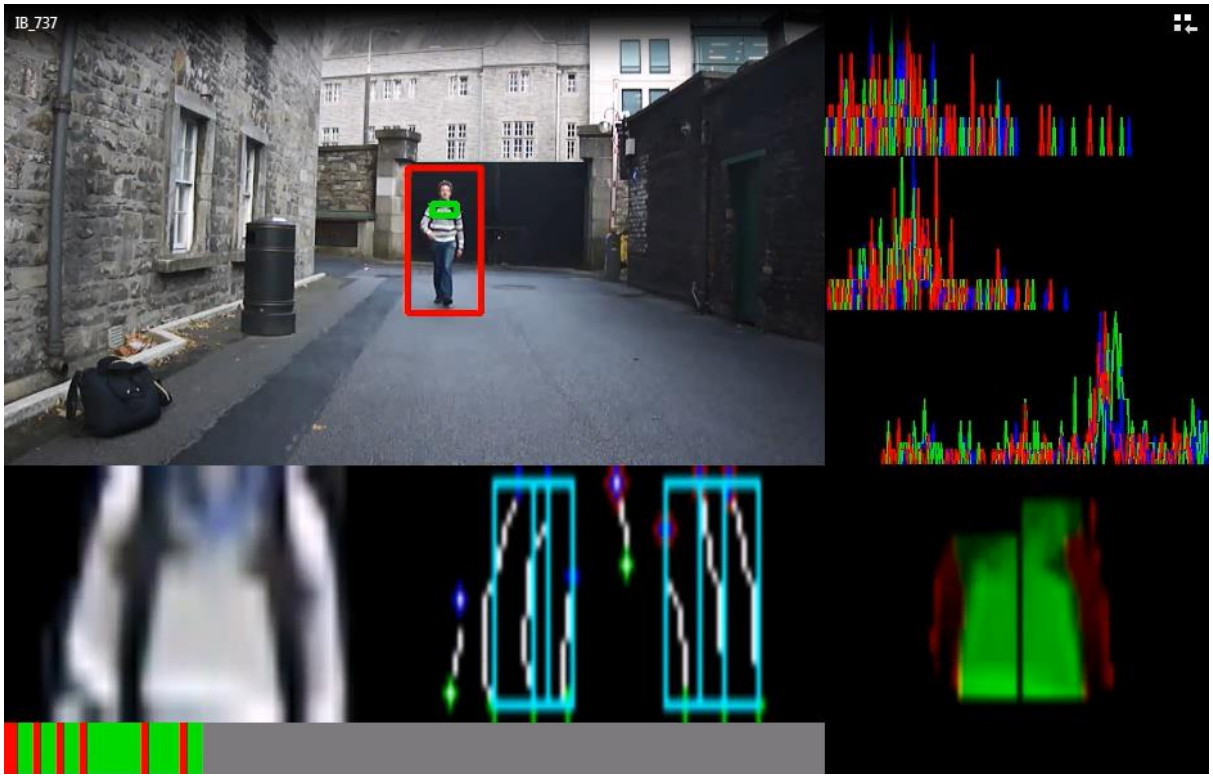


Figure 49: Backpack being detected

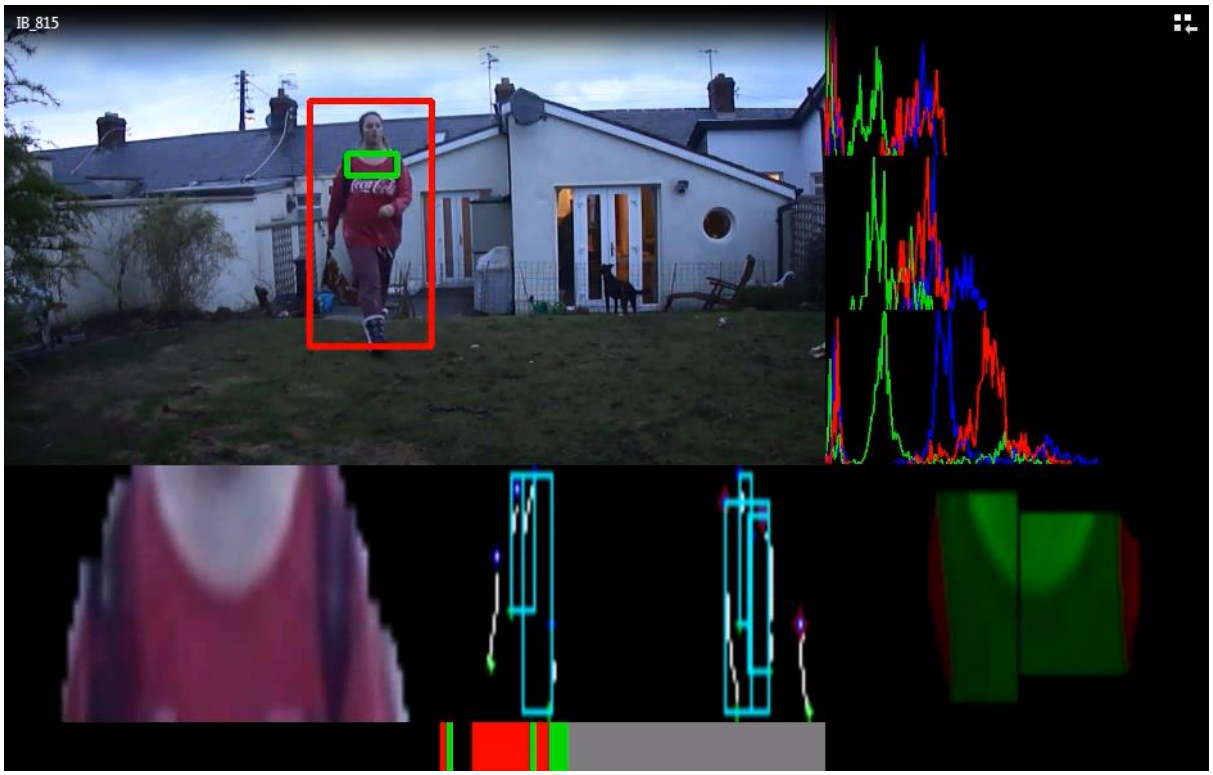


Figure 50: Backpack being detected

Figure 51: No backpack detected and none present

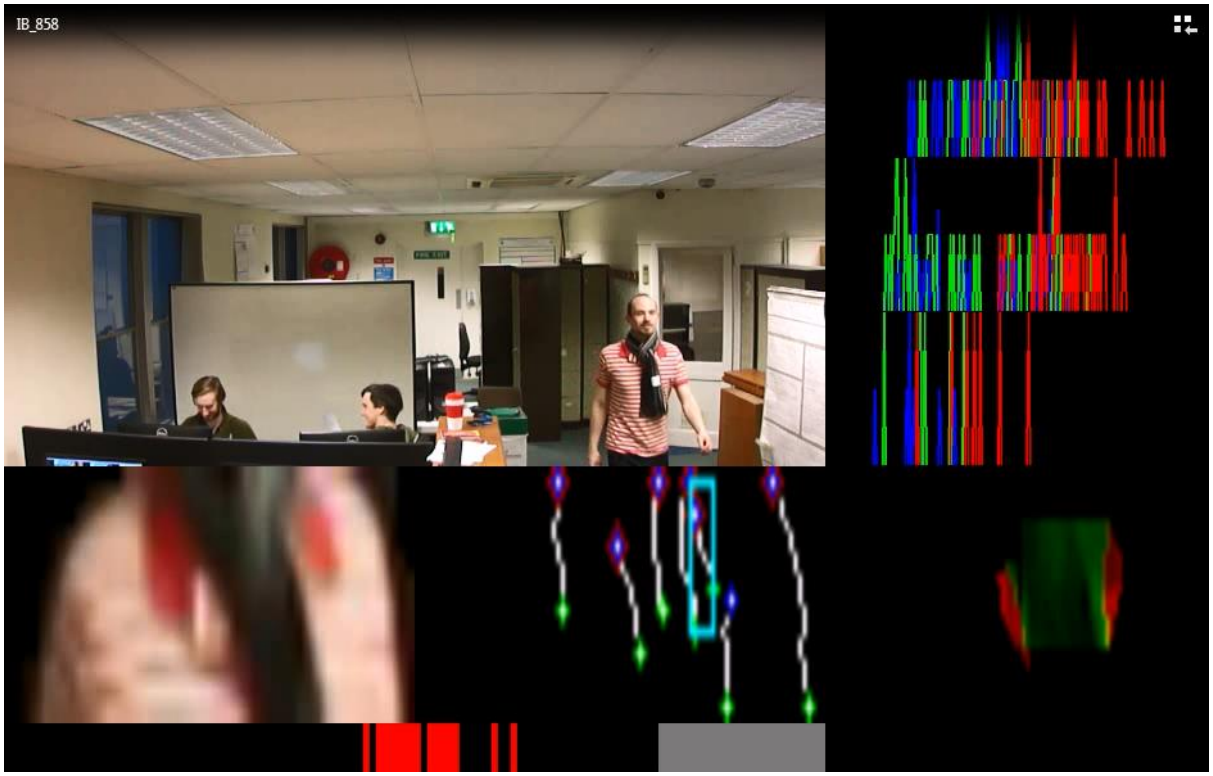


Figure 52: No backpack detected and none present

8.5.2 Failures

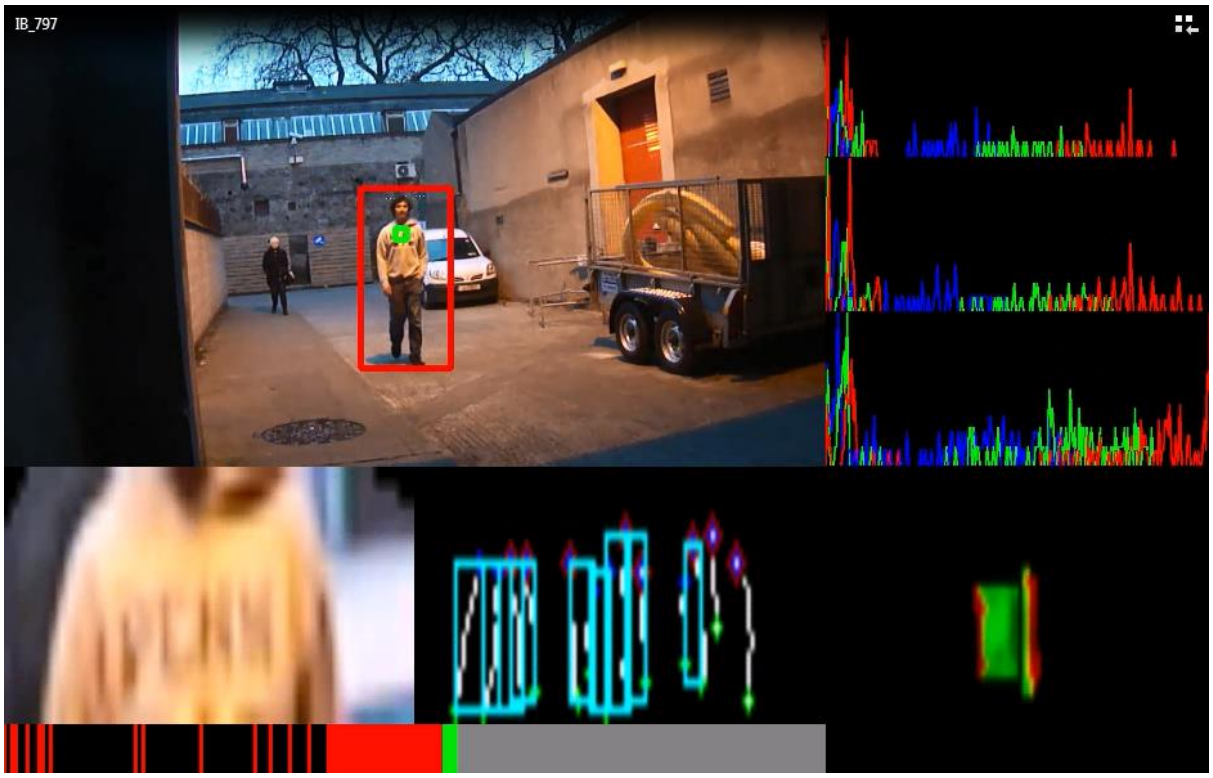


Figure 53: Logo causing false detection

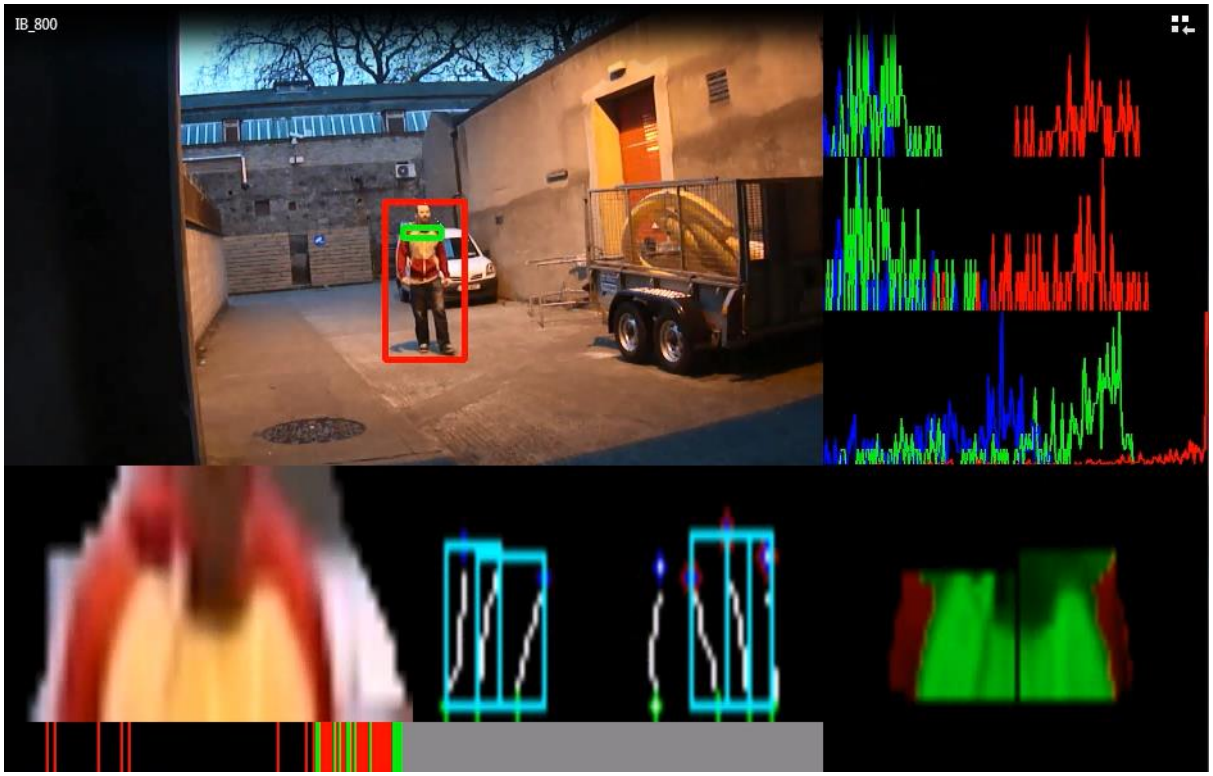


Figure 54: Red garment and background subtraction failure causing red patches at edge of garment to be considered as straps

This method is relatively good at detecting backpacks when they are visible within a scene as can be seen in *fig. 49 & 50*. It is relatively easy to isolate potential straps and classify them as being actual backpack straps. This system is also much better at not detecting backpack's when they are not present as is visible in *fig. 51 & 52.*

However this method can still fail when there are regions of similar colour and shaped like straps present in an image. As can be seen in *fig. 53,* the red regions of the garment have passed the colour histogram check as they have a different colour to the inside garment. These potential straps should have been removed before the method even reached the stage of colour checking, as some of the detected edges are too close to the edge of the upper torso region. However, the background subtractor has included parts of the area around the foreground mask as foreground. Hence these edges are far enough away from the edge to not be excluded by background subtraction.

An additional failure can be seen in *fig. 54* where the logo regions are being classified as potential straps as they consist of letters with long vertical arms of a different colour to the underlying garment.

The combination of looking for parallel edges and combining this process with another method to increase the accuracy has enabled the lower of the Sobel based edge detector threshold by a significant margin. It has been reduced from an optimal of 134 (given a pixel range of 256) for approach three, to an optimal of 40 for approach four. This ensures that more edges are picked up and analysed leading to a reduction in the number of false negatives. Addition false positives created due to the lower threshold are caught by the colour histogram checking method.

There are several weaknesses in this system that could potentially eliminated by further improvements. For instance, there are several cases, where a background subtraction failure has left regions at the edge of the real upper torso region being considered as straps. A potential solution to this problem would be to segment out the garment region between the outside of the strap and the edge of the upper torso region. This could then be compared using colour histograms to ensure it matched the garment region between the straps. An additional check would be to ensure that it was also of a different colour composition to the straps themselves. This would catch cases as shown in *fig. 54.* However, care would have to be taken with such a method, as any inclusion of the background within the foreground region would result in the outer garment region having a different colour composition when compared to the inner garment region. This could fail true positives that would otherwise have passed the detection criteria.

Additionally, statistical means could be introduced to provide additional checks upon strap location, by ensuring they progressed over the tops of shoulders and so on. The symmetry checks used in approach three, that were removed from approach four, could be re-introduced, to eliminate cases such as those visible in *fig. 53* that eliminate logos from detection.

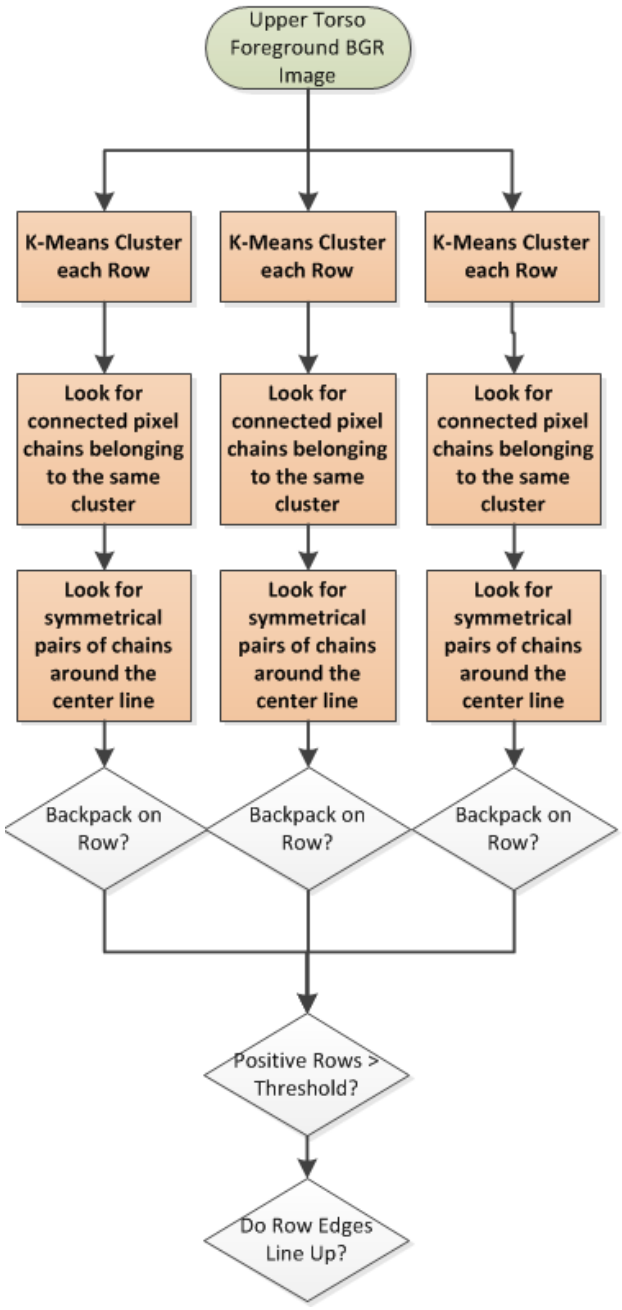
8.6 Summary

Approach four has the highest Matthews and accuracy results seen by any approach yet. This demonstrates that checking the result of approach three with another method enables greater detection precision. This suggests that combining methods enables the elimination of weakness in one method.

9 Approach Five: Row by Row Colour Space Clustering

This chapter examines the fifth and most successful approach. This takes the colour clustering techniques discussed in chapter 5 and applies them to every row instead of the whole image. Needless to say the statistical analysis applied to each is very different to that used in the first approach. The reader should read section 5.2 of approach one to obtain an understanding of the K-Means clustering algorithm used in this chapter.

The rest of this chapter contains a high level overview of the design in section 9.2. The analysis used to determine if each row matches the conditions for a backpack, is in section 9.3. The combination of information from all of the rows is discussed in section 9.4. Section 9.5 looks at the additional checks that were applied to the image to further enhance the accuracy of the detection method. The results are presented in section 9.6, analysed in section 9.7 before a summary in section 9.8.



9.1 Design Overview

1. **Row Analysis** – For every row within the region of interest we applied the background subtraction mask so that we only had pixels relating to the foreground object.
2. **Colour Clustering** – The pixels of this row were clustered depending upon their RGB values using K-Means clustering.
3. **Row Classification** – This row was then analysed looking for connected chains of pixels within the same cluster. The centre point, length and left/right boundaries of these chains are recorded.
4. **Symmetry Analysis** – Each of these rows is analysed, looking for pairs of chains with similar length positioned symmetrically a similar distance from the centre line of the upper torso. If such a pair of chains is found, this row is returned as having a potential backpack present and the bounds of the strap are recorded.
5. **Percentage of Rows** – When all rows have been individually analysed, a check is made to see if an adequate number of them have returned a candidate for a backpack. If the localised

percentage of rows with a potential backpack is above a threshold percentage we continue analysing.

6. **Row Continuity** – For each row from the top to the bottom we computer the difference in horizontal location between the centre points, left boundary and right boundary of each strap. These differences are all added together and combined to produce an overall fit value for the whole image. If this fit value is above a pre-defined threshold the frame is rejected.
7. **Strap Width Variation** – The standard deviation of the strap widths over all of the rows is computed to ensure it is below a certain threshold. If it is we consider this frame to have a backpack present.

9.2 Clustering Individual Rows

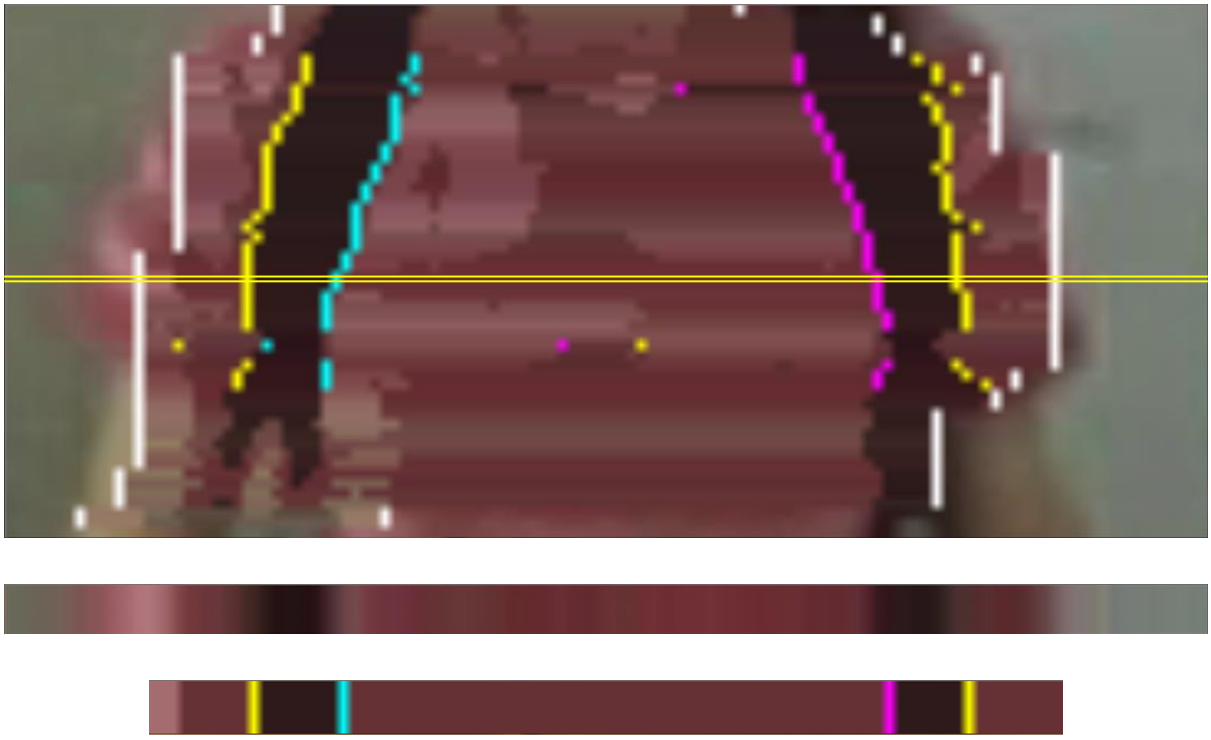


Figure 55: Clustered Image with selected row bounded by yellow bars (top). Un-clustered and enlarged image of that row (middle). Row after K-Means clustering has been applied and truncated to only the region within the UTR (bottom).

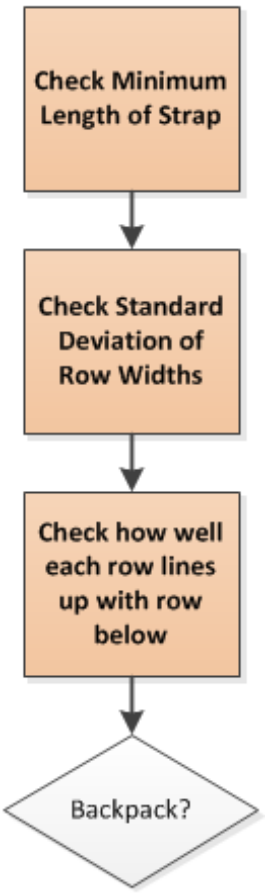
Backpack straps present long vertical edges that are more easily detected by analysing the image in the horizontal direction than the vertical direction. Hence, by K-Means clustering the image along rows as opposed to the whole image, the boundaries of the strap will be picked up by the clustering process. If we observe *fig. 55* which demonstrates clustering over the whole image, we can see how the strap's edges are poorly detected as opposed to *fig. 55* where the strap boundaries have been picked up much more clearly. The ideal number of clusters was found to be 3 after parameter tuning, as this roughly left one cluster for the strap, one for the background region and another for portions of the image that don't fit the other clusters. 2 clusters increased the chance of straps and the underlying garment being assigned to the same region. 4 or more clusters did not provide any improvement and increased the probability of each strap being assigned to different clusters or

slightly darker regions of the underlying garment being in-correctly classified as straps as shown in *fig. 53*.

9.3 Statistical Analysis of Individual Rows

As we are looking for double strap backpacks we want to find two regions in each row that both belong to the same cluster. We want both of these to be within a similar length of each other as a large asymmetry in size suggests that they are not from the same object**.** In addition pairs of backpack straps are usually found in symmetry with each other, the same distance from the torso centreline. Hence we will only look for straps that are symmetrical. If two regions that satisfy this requirement can be found in a row it will be labelled as having a potential backpack. In *fig. 53* all the rows that meet this requirement have had the outer bound of each strap coloured yellow and the inner bound coloured cyan for the left strap and magenta for the right strap. The middle and bottom rows show the process of clustering being applied to a single row and shows the two black regions identified as being symmetrical in width and distance from the centre line.

9.4 Statisitcal Analysis of All Rows



As can be seen in *fig. 56* there will be a certain amount of incorrectly classified rows as indicated by the lone cyan, magenta and yellow dots that do not align to the strap boundaries. However as long as there are a great enough number of rows within the upper torso region these errors can be ignored as they will not affect the result of the overall classification.

Additional checks are applied to reduce the number of false detections. The system finds the uppermost row that has returned a backpack and the lowermost row that has returned a backpack. This gives the length of the backpack straps which is compared to the pre-defined minimum length. This check was introduced to prevent lapels, *fig. 56,* shirt pockets and other items from triggering false detections. A backpack strap will

Figure 56: False detection due to short straps (top). False detection due to un-aligned regions (upper middle). False detection due to widening scarf (low middle).

usually be detected over most of the upper torso region and hence will have a greater length.

9.5 Additional Statistical Checks

As indicated in *Fig. 56* there were occasion when many rows would returned potential candidates for backpacks that did not line up with other rows. These were clearly not backpacks however they resulted in false detections. The easiest way to rid the system of these errors was to check how the boundaries produced by each row lined up with each other. This was done by simply adding the horizontal distance between each boundary on one row with its position on the next row. These distances were averaged over all of the rows. If the position of the boundaries varied by a large amount as shown in *fig. 56* the produced total would be rather large and the straps eliminated.

A second group of failures was caused by other regions being evaluated as straps such as the top of the scarf as indicated in *fig. 56*. Backpack straps tend to have roughly similar width over their entire length; hence we can discard items with a varying width over the length. The standard deviation of backpack strap width is calculated over all of the rows and compare to a threshold for this purpose.

9.6 Results

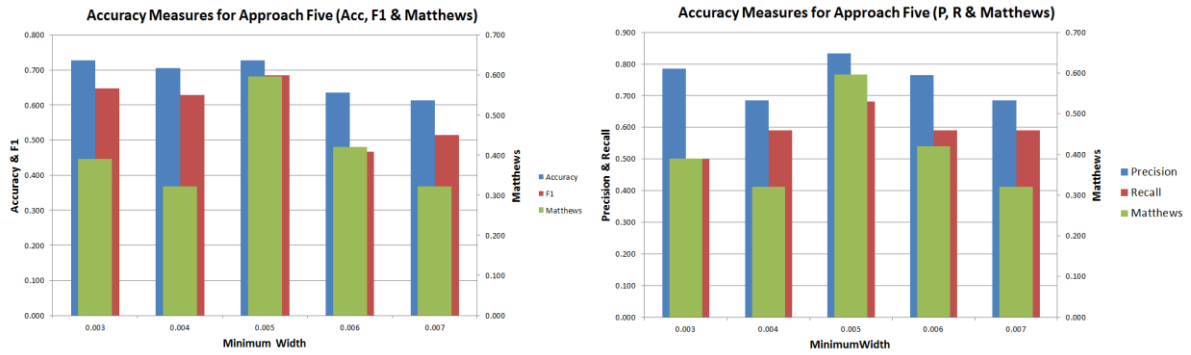
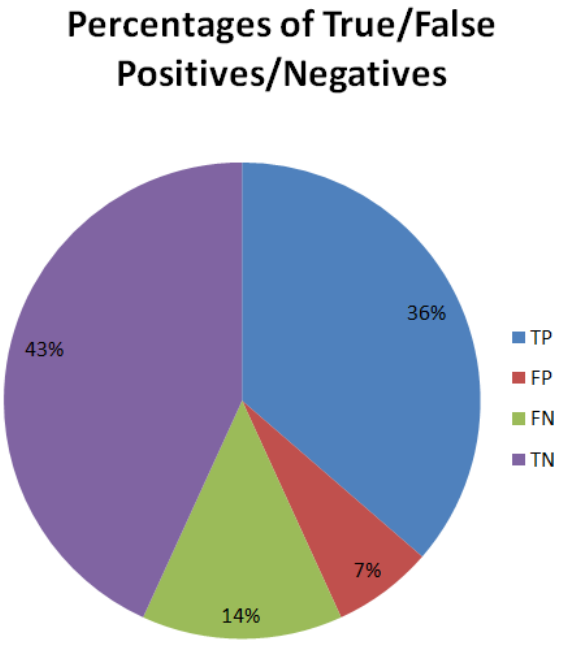


Figure 57: Graphs for Approach Five

Approach Five gives the best results of any of the approaches with a high level of true detections and a low number of false detections. There were 6 tuneable parameters used in this approach:

- **ROW_WIDTH_VARIATION** – This governs the allowed amount of variation in the width between two chains being considered as potential straps on a row. It is a value given relative to the width of the longer potential strap. During the process of tuning, the optimal value was found to be 0.5.
- **THRESH_NUM_ROWS** – This governs the minimum percentage of rows that must indicate a potential backpack strap along the length of the suspected strap. During tuning the optimal value was found to be quite low at 40%.



Asymmetry	TP	FP	FN	TN	Recall	Precision	Accuracy	F1	Matthews
0.2	16	3	6	19	0.457	0.842	0.795	0.593	0.596
0.1	11	1	11	21	0.344	0.917	0.727	0.500	0.510

Figure 58: Best results for Approach Five

- **MIN_WIDTH** – Defines the minimum width that a potential strap can have before it is not considered. The value is given relative to the width of the upper torso region. During tuning this was found to be an un-necessary hindrance to the system and ended up being set close to 0 at 0.005.
- **STRAP_SYMMETRY_VARIATION** – Optimally two straps will be positioned symmetrically around the centre line. Inevitably there will be some variation to this ideal and this parameter controls the maximum allowed variation between the two potential straps distance from the centre line before they are considered to not be straps. This value is given relative to the width of the upper torso region and during tuning it was found that the optimum value was 0.2.
- **ROW_LINE_UP** – This threshold governs the level to which the strap boundaries on each row must line up with each other. During tuning it was found that the optimal value was 8.0.
- **SD_VAR** – This is the allowed standard deviation of the strap width for all rows that report a potential strap. Through tuning it was found that the optimum maximum value was 120.

9.6.1 Successes and Failures

35 of the videos ran successfully, there were 3 cases of false positives caused by:

- Jacket lapels being detected as straps. A background subtraction failure prevented rejection due to their position at the edge of the upper torso region *fig 60 top*.
- Small Boundary region between scarf and underlying garment due to the tuning process setting the minimum width of a strap to be small**.**
- Shadows occurring across a two toned garment causing clustering to create two strap like regions as shown in *fig 60 left*.

There were 6 false negatives caused by:

- In 2 cases not enough individual frames were detected.
- In 2 cases a scarf worn by a test subject caused occlusion of the underlying straps as shown in *fig. 60 right***.**
- In 1 case the arms were included in the background region, causing the straps to be located at the edge of the foreground region and hence classified as not straps.
- In 1 case a two-toned garment prevented the straps from being clustered correctly.

There were 35 successful videos as seen below.

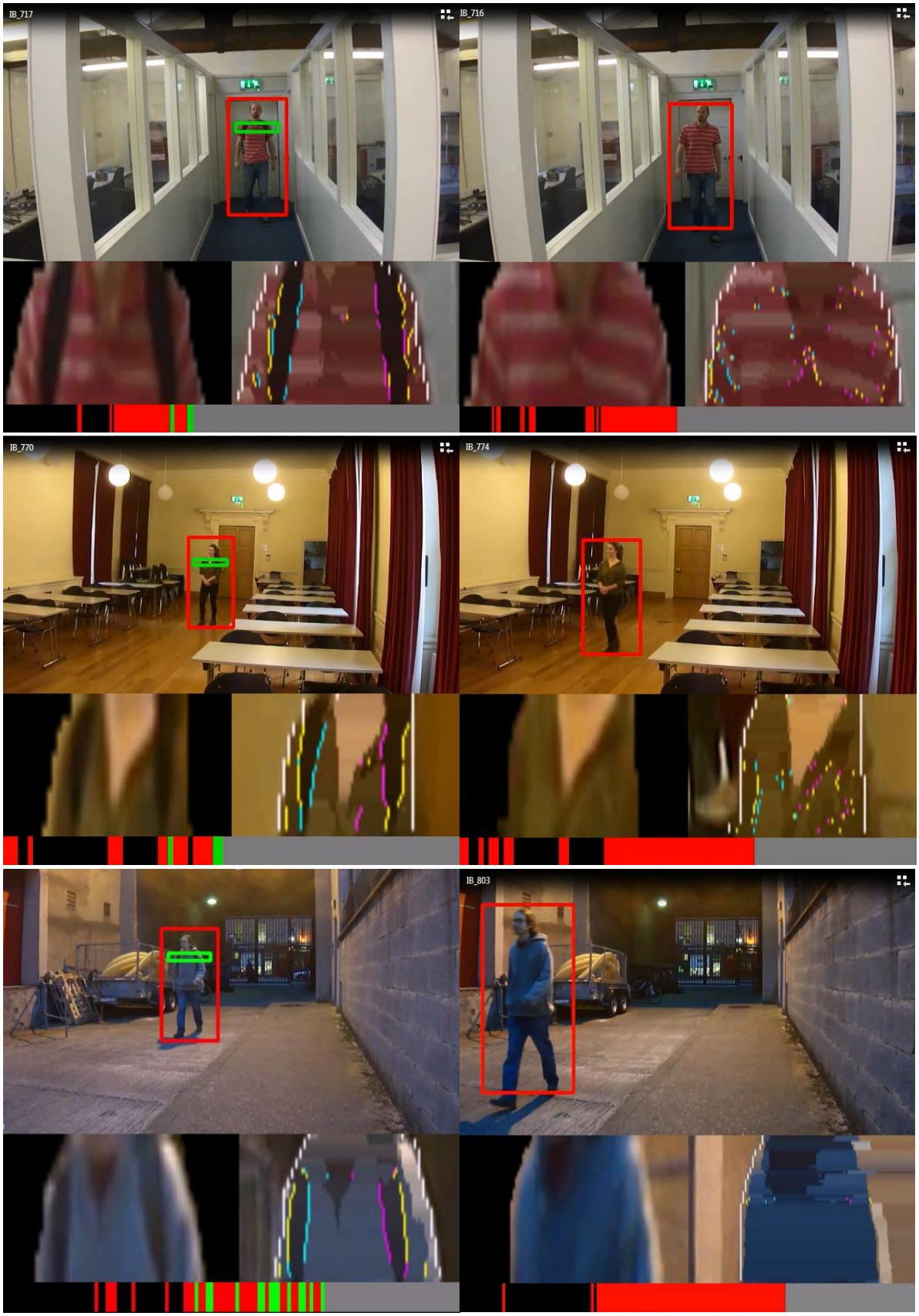


Figure 59: Successes

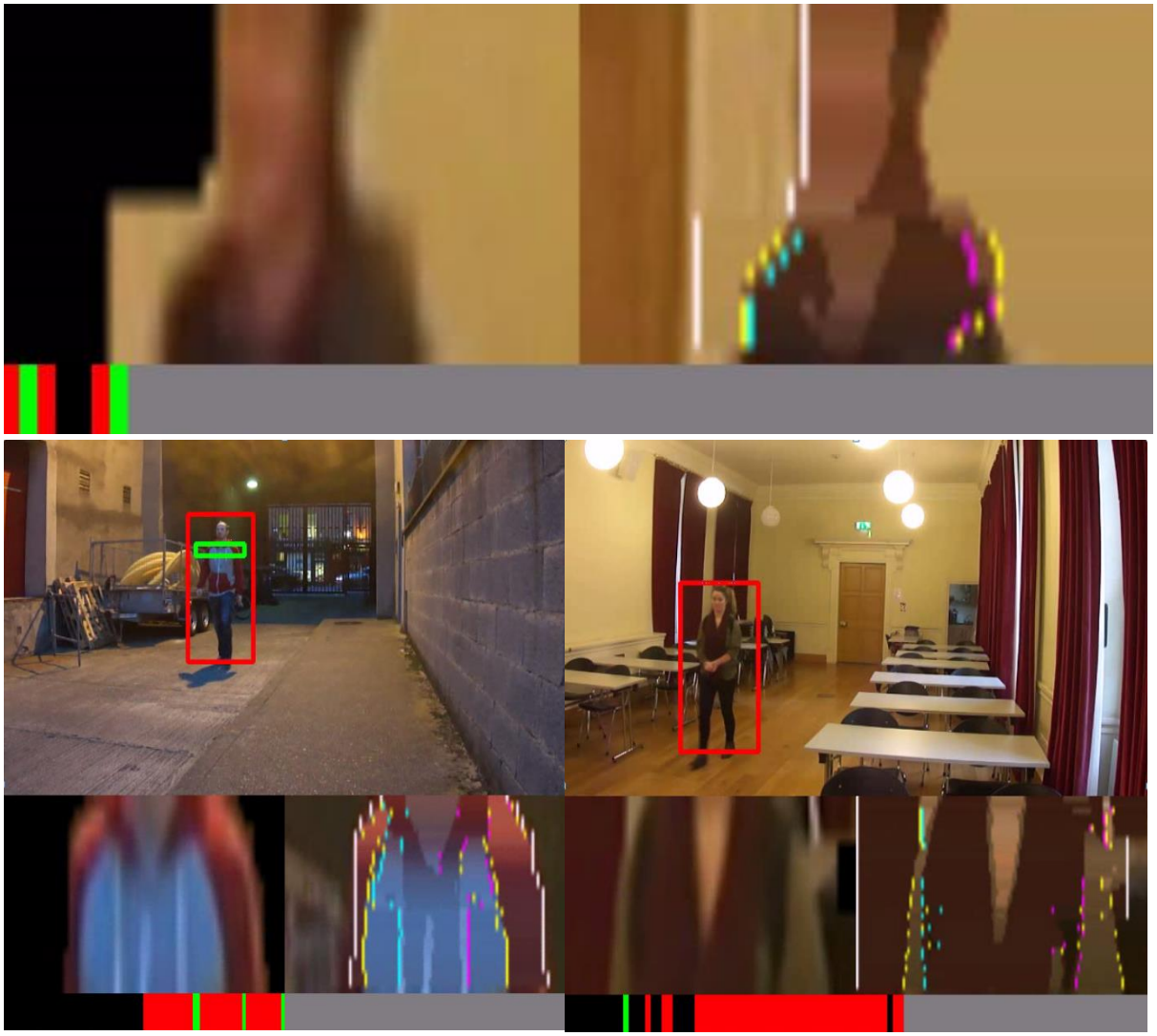


Figure 60: Failures

9.7 Evaluation

The false positive rate is very low for this method with a value of 0.136. The system was tuned until a maximal value of the Matthews correlation co-efficient was obtained. However another run was obtained that had an even lower false positive rate of only 0.045. The parameters used in this run traded a small reduction in the number of detections for increased precision within those detections. Depending upon the requirements of the system a lower false positive rate may be more desirable, particularly if the false negative rate is not as important. However in other situations the reverse may be true, for instance if the system is highlighting footage for an operator to review a false positive will be recognised by the operator however they will never view the false negative case.

The false positive caused by the scarf is quite annoying to the author as extra code was specifically written to eliminate potential straps whose width varied too much along their length. The scarf widens towards the end which would trigger this code causing a correct true negative. However as the minimum width of a strap evaluated as very low value during parameter tuning the boundary

between scarf and underlying garment is now being detected as a strap. Increasing the minimum width could eliminate this failure however it will also re-introduce failures as it takes the parameters away from their optimum setting. To eliminate this failure would require looking down the minimum width parameter and re-tuning the system to optimise its performance.

The false positive visible in *fig 60 left* is notoriously hard to eliminate and has caused numerous failures during the system development relating to the two toggles and zipper which are all assigned to their own cluster in many rows. The boundaries caused by the zipper, two toggles and red patches of the garment, perfectly segment the white region into two strap proportioned regions. It is very hard for the detection system to correctly eliminate these. Potentially a colour histogram check, as used by approach four, could be re-implemented as an add on to this approach to check for cases, where a shadow accidently classifies the main region of the garment as two strap regions. However even this may not work for this particular video as the red garment may confuse the system into thinking the white of the fake straps was in fact, actually a different garment type.

In terms of the false negatives:

In the two cases shown in *fig. 60* occlusion by scarfs is causing the K-Means clustering to fail to correctly cluster the straps. The first case is going to always be hard to eliminate if not impossible, as the scarf is covering the straps themselves. Hence we cannot expect the system to correctly classify cases such as this one. The second case could potentially be eliminated if the system was designed to, not take into account the centre region of the torso. This would work for double strap backpacks; however, it would introduce problems when trying to detect singe strap bags that go across the torso.

In *fig. 60* the white lines indicate the edges of the upper torso region detected by the system. As can be seen these exclude the arms from consideration placing the straps at the edge of the detected region. Hence the straps are eliminated from contention for detection as a strap. This failure is caused by a background subtraction failure due to the similarity between the white garment and the white walls behind the object. This failure could potentially be eliminated with further tweaking of the background model or a more advanced background subtraction model. Neither of these items are the primary focus of this thesis.

The last false negative is caused by the two-tone garment ironically the same two-one garment caused a false positive when viewed on its own without a backpack. In this case the darker colour is blending in with the backpacks during the clustering process. This results in the straps not being detected at the optimal locations present at the top of the upper torso region.

9.8 Summary

In my view this approach works very well and has the ability to detect straps with a low level of contrast relative to the underlying garment. Most of the failure cases are related to pre-processing steps or are very difficult to solve even using other approaches.

Potential improvements to the this approach would be to incorporate a colour histogram check of detected straps, as detailed in approach five, to try and eliminate some of the false positives. Currently this approach used absolute location checks, for instance, a potential strap is classified as being too close to the edge of the upper torso region to be a strap or not. This could be replaced with a weighting system that would allow out of position straps to be classified as backpacks if they satisfied other criteria to a significant enough extent.

10 Approach Six: Single Strap Detection

As approach five achieved the best results, it was used as the basis for a single strap detection approach. This was a very basic modification with parts of the algorithm relating to finding two straps removed. Instead of running the whole solution over the entire upper torso region it is called twice, once for each half of the region. Only one of these needs to find a strap for the upper torso region to be classified as containing a single strap bag. This chapter will open with a high level design overview in section 10.1, with the differences relative to approach five detailed in section 10.3, before moving onto results in section 10.2 and evaluating them in section 10.3. It is assumed the reader will have read approach five for comprehensive understanding of this chapter.

10.1 Design Overview

1. **Row Analysis** – For every half row within the region of interest, we applied the background subtraction mask so that we only had pixels relating to the foreground object.
2. **Colour Clustering** – The pixels of this half row were clustered depending upon their RGB values using K-Means clustering.
3. **Half Row Classification** – Each half row was then searched for a connected chain of pixels that satisfied the requirements for minimum and maximum width of a strap.
4. **Position Analysis** – This connected chain was analysed to ensure that it was positioned roughly in the centre of the half region, as backpack straps do not usually reside at the edge of a torso or in the middle of the torso.
5. **Percentage of Rows** – When all rows have been individually analysed we check to see if an adequate amount of them have returned a candidate for a backpack. If the local percentage of rows with a potential backpack is above a threshold percentage we continue analysing.
6. **Row Continuity** – For each row from the top to the bottom we computer the difference in horizontal location between the centre points, left boundary and right boundary of each strap. These differences are all added together and combined to produce an overall fit value for the whole image. If this fit value is above a threshold the frame is rejected.
7. **Strap Width Variation** – The standard deviation of the strap widths over all of the rows is computed to ensure it is below a certain threshold. If it is we consider this frame to have a backpack present.

10.2 Differences from Approach Five

Most of this approach is identical to approach five; however, we can no longer rely on the symmetry of two straps to use as an eliminating factor. Instead this approach places greater emphases on the minimum and maximum width of potential straps. It also relies on the position of the strap within the upper torso region to a greater degree than approach five.

10.3 Results

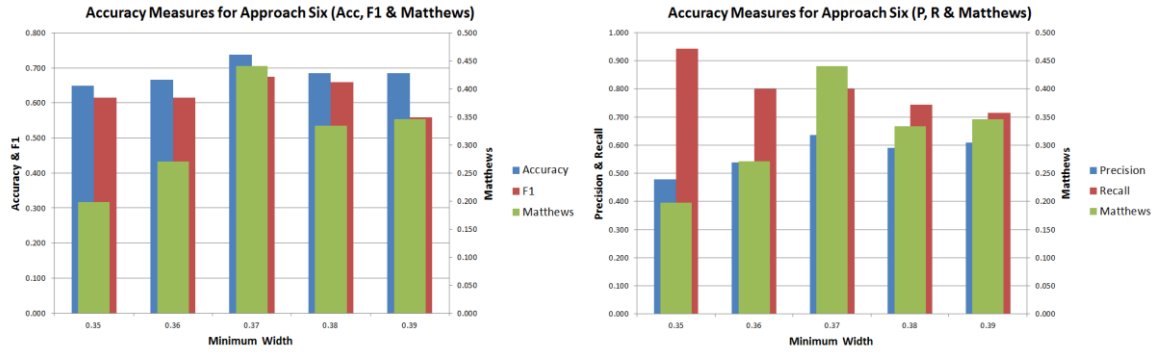
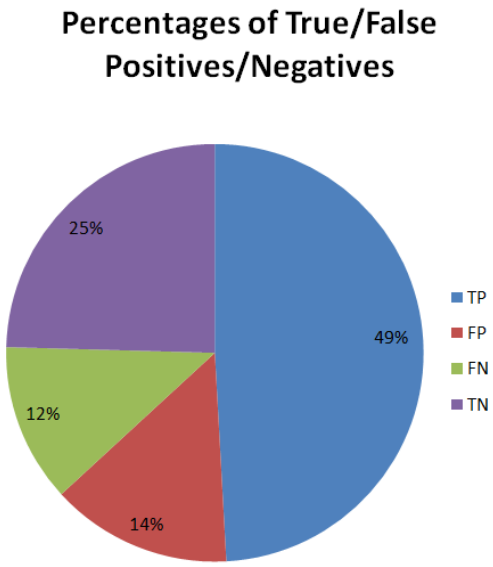


Figure 61: Minimum width being tuned for Approach Six



TP	FP	FN	TN	Recall	Precision	Accuracy	F1	Matthews
19	8	16	14	0.576	0.704	0.579	0.633	0.175

Figure 62: Best Results for Approach Six

Approach Six was developed late in the development cycle and hence, the time could not be afforded to tune it. Hence, these results were obtained using the parameter values from approach five since they shared many of them. Performance would be significantly improved if the parameters could be tuned specifically for this approach.

The parameter that could be tuned as the minimum width of the strap as is shown in the graphs presented in *Figure 61*.

10.3.1 Successes and Failures

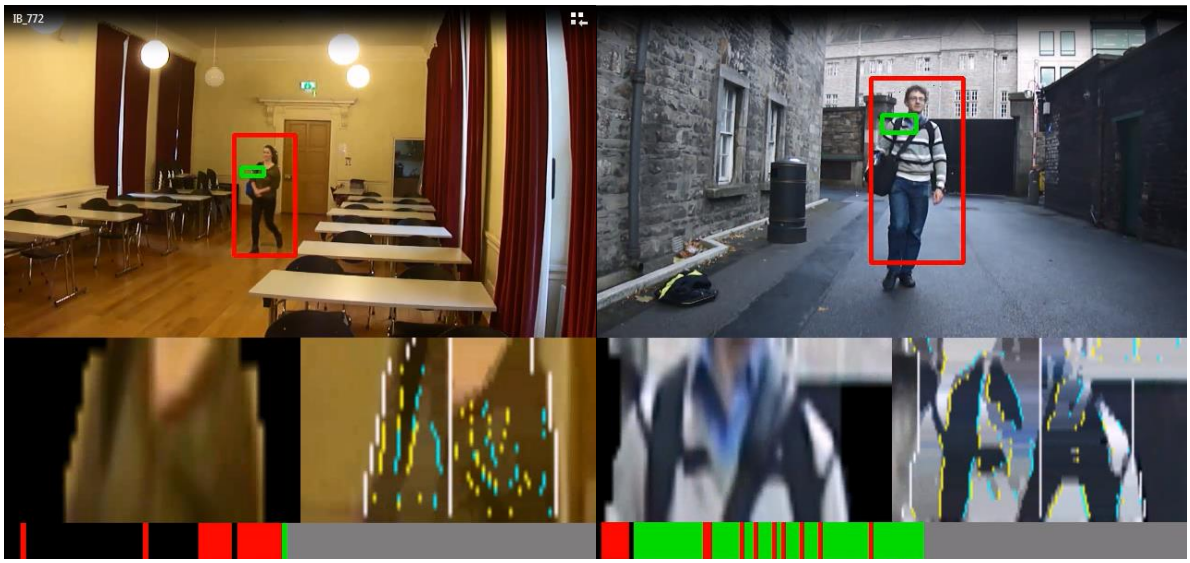


Figure 63: Successful detection of a single strap (left) and a single strap of a double strap backpack (right)

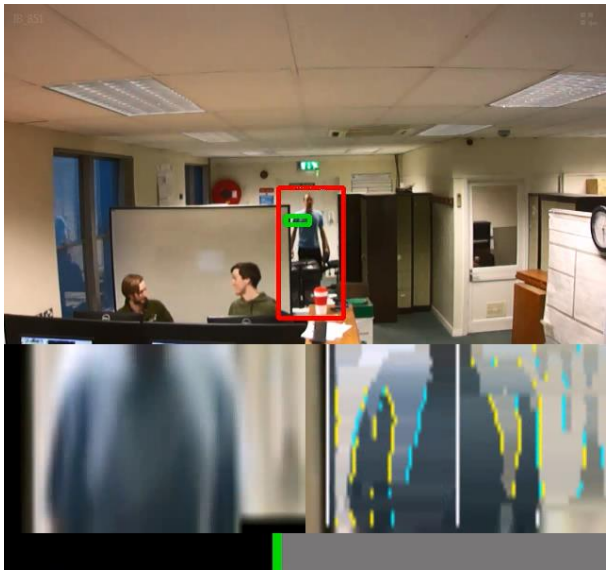


Figure 64: False detection caused by strap like regions created during the clustering process. This was introduced by the bad illumination conditions present in the room. White line indicates the divide between left and right clustering region.

for approach five.

10.4 Evaluation

Considering how little modifications were made to approach five to enable it to detect single strap backpacks the performance is surprisingly good. The lower level of accuracy that was present was because the system could not be tuned as much as the other systems due to the limited time available and the decision to priorities the double strap methods. This could probably be restored to at least the accuracy of approach five if the time needed was available.

The types of failure are much more varied as there is not second strap to perform a symmetry check against. Hence false detections are much higher for this method than they are

10.5 Summary

This approach has shown that it is possible to re-apply the double strap methods employed in previous approaches to the task of detecting single straps. While there is a performance drop this is more likely due to the low amount of tuning that was applied to the system due to time constraints.

11 Conclusion and Future Work

11.1 Introduction

At the outset of this project the goal was to investigate and test methods to detect backpacks from frontal views of individuals in surveillance videos. This was a challenging task as at the beginning of the project there was no published literature available on the problem nor did publically available third party test data exist. Hence the test data and methods used had to be developed from scratch. The testing portion of the goal has been achieved through the creation of a representative test set that enables benchmarking of systems. The goal of detecting individuals has also been successfully achieved with six approaches of varying accuracy produced. The results obtained for these six approaches are summarised and compared in section 11.2. From this the overall conclusions of the project are drawn and discussed in section 11.3. Potential areas that this project can be continued on from are outlined in section 11.4.

11.2 Comparison of Approaches

Approach	TP	FP	FN	TN	Recall	Precision	Accuracy	F1	Matthews
1	7	6	15	16	0.318	0.538	0.523	0.400	0.050
2	16	10	6	12	0.727	0.615	0.636	0.667	0.277
3	17	9	5	13	0.773	0.654	0.682	0.708	0.370
4	18	9	4	13	0.818	0.667	0.705	0.735	0.420
5	16	3	6	19	0.727	0.842	0.795	0.780	0.596
6	28	8	7	14	0.800	0.778	0.737	0.789	0.440

Figure 65: The best results obtained for each approach after tuning.

This section summaries and compares the results of the six approaches developed to find complimentary solutions. First off it should be noted that approach six designed to only detect single straps while the other five methods were all designed to detect double strap backpacks. Hence the test data used by approach six is slightly different to the other five methods. This means that the figures available for approach six can't be directly compared to the other five approaches. In addition as this was the last approach developed it was not as well tuned as the other systems. Therefore the following analysis will concentrate mainly on the first five approaches.

The initial approaches for both colour and gradient analysis (approaches one and two respectively) have the poorest results. Not surprising the best results came from the later developments of these methods (approaches four and five). Colour space analysis produced better results than gradient based edge analysis. This was due to the ability of the colour based approach to detect straps with a low level of contrast relative to the underlying garment. Additionally small edges created by jacket lapels, pockets and logos and other items were more likely to confuse the edge analysis. The row based colour space analysis was much better equipped to discard these items as they would not appear on enough rows to be considered as a strap.

All three of the gradient based methods had higher levels of recall than precision. This was reversed when the two colour based methods are analysed as seen in *fig. 66*. This suggests that gradient analysis methods developed in this dissertation are more likely to produce false positives while colour analysis methods are more likely to produce false negatives. Hence it suggests that the gradient and colour analysis methods are complimentary to each other and could be used to re-enforce each other's results and improve overall accuracy. A basic implementation of this was already used in approach four which had an underlying edge based detection method tuned to be extra sensitive. A colour histogram check was then employed to confirm or reject these results leading to an improvement in the systems accuracy. It would be a useful exercise to repeat this process with a colour clustering technique as well as a colour histogram technique to see if further improvements can be achieved.

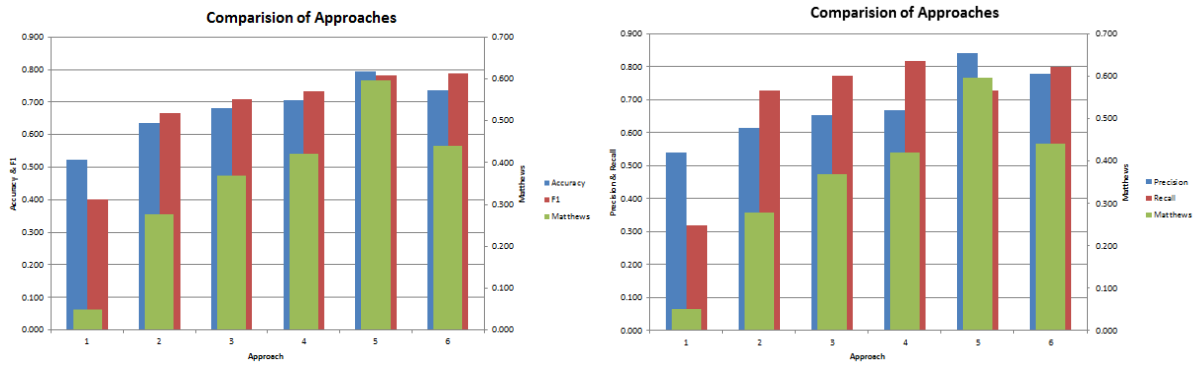


Figure 66: Precision, Recall, Accuracy, F1 and Matthews values for the best results of all six approaches.

11.3 Conclusion

The main goal of this project has been met as a system has been developed that can detect backpacks on individuals in surveillance footage. Currently surveillance operators can be overwhelmed by the hours of footage they have to view. This can be drastically reduced by a system such as this which aids by highlighting footage that contains items of interest such as backpacks. This system looked to achieve this by searching for straps in front views of the individual, an area not well covered in current literature. There is no published work available on the subject of detecting backpacks from a front view of an individual in outdoor environment, a situation covered by this project.

As part of this dissertation a test database was created that will fill a gap that is currently present in publically available test data. This was built with the aim of being publically releasable to enable verification of the results for this dissertation. Hence it will hopefully represent a contribution to the available test data and can be used to benchmark other frontal backpack detection systems. It should be noted that the only paper that has addressed the issue of frontal backpack detection did not make its test data set publically available.

The system, in particular solution five, is successful at detecting straps that have a low contrast level relative to the underlying garment. This is superior to the performance given in [30] where a high level of contrast was assumed. The authors of that publication also made the assumption that the

strap would be darker than the underlying garment. All of the solutions here will function equally well with either the straps or underlying garment being darker relative to the other. In addition the test data of that publication was taken from inside a brightly lit airport. The test data created as part of this project is a combination of indoor and outdoor footage in varying conditions. As can be seen in *fig 59* the system managed to perform in twilight conditions under a street lamp.

The system has two main limiting factors; the first is that none of the solutions developed can cope with situations where there is no contrast between the straps and underlying garment. This issue will be very difficult to solve using computer vision as without contrast even the human eye cannot distinguish the strap from the underlying garment as shown in *fig 59***.** Generally computer vision systems are considered unable to detect items that cannot be distinguished by the human eye [36]. Unfortunately this situation was found to be quite common when the author analysed garment and backpack combinations worn by pedestrians entering Trinity College during two February mornings.

The second limiting factor of the system is that it cannot cope with occlusion of the backpack straps. This is also a difficult problem to solve as there are many different ways the straps can be occluded either partially or fully. A problem such as this would require several techniques to solve and once again runs up against the general rule that the straps can only be detected as long as they are visible to the human eye.

Despite these deficiencies in the system I believe that it can be combined with the already available solutions for detecting backpacks from the sides [19] and back [30]. This would produce a system that is able to detect humanly visible backpacks on pedestrians from several angles. This would be of benefit, helping security personnel to narrow down the hours of footage they have to review. It should be noted that as with any binary classification there is still some levels of inaccuracy present in the system. As has been noted in the evaluation sections for each approach this is going to be difficult to totally eliminate as scenarios can always be encountered that will trick the system. The next section details improvements that can be made to the system to increase its reliability and improve its detection rate.

To put the results achieved in this project in perspective keep in mind that frontal backpack detection has not been addressed outdoors or in variable illumination conditions to date. Due to the limitations of computer vision it is also not possible to detected cases that can't be easily distinguished by the human eye. For instance black straps on black jackets will be nearly impossible to detect. Detection of these items will require an alternative method such as thermal spectrum analysis as advised in the future work section. Other publications have achieved similar levels of accuracy such as [26] which achieved a precision rate of 50.5% and a recall rate of 55.4% when trying to detect carried items on individuals in a train station in Leeds.

Computer vision systems barely ever consist of just one stand-alone component. The solution presented in this paper would be merely one part of an overall backpack detector that detects from several angles using several methods to ensure reliability.

11.4 Future Work

As solutions four and five had the highest detection rates, future improvements to the system should be applied to these methods. There is merit in keeping both of them, as they analyse video footage in gradient space and colour space respectfully complimenting each other. There are several cases where a false detection in one is correct in the other. A potential area for further investigation would be to work out how to combine these two methods so that when they disagree the classification is made using the system that performs best given the local frame conditions. Approach four made use of this technique however only with colour histograms. It would be interesting to integrate otherwise fully stand-alone gradient and colour based approaches.

As mentioned in section 3.4 the test data set is relatively small and hence may not test the system as comprehensively as would otherwise be possible. Future research should look to expand this data set and analyse the performance of the current system. With more data the cases that cause the most false detections can be identified. Additional statistical checks can they be added to the system to counteract these failures and improve the robust and reliable behaviour of the system. Additionally if multiple parties contribute to the data set it should provide a greater variety of scenes and clips which will increase the effectiveness of the data set.

A larger data set will also enable more reliable tuning of the system. The tuning employed in this project involved manually comparing the results obtained from several automated runs with different parameter values to achieve the highest Matthews score. This process could be automated to reduce the time spent on it and gain better results for the system.

As detailed in chapter 10 the system has the ability to detect a backpack if only a single strap is worn. However this was a quick modification made to the two strap version of approach five and applied twice on both sides of the image. An improvement to this would be a re-implementation of approach five that only needs to be applied once over the whole upper torso region and can detect both double and single straps. Both single straps that cross the upper torso region to go over a shoulder opposite the bag and vertical single straps such as handbags need to be considered in this approach. As approach four detects single straps while searching for pairs of parallel contours an approach

The main limitation of the system is its inability to detect straps that do not contrast with the underlying garment, an issue that is hard to solve in the conventional BGR colour space. A potential solution to this problem would be to detect the presence of backpacks using a thermal imaging camera. As the strap will be made of a different material than the underlying garment it will have different thermal properties and emit heat at a different rate. Approach five uses colour classification and may not work as well in this instance. Approach four is based upon edge detection and should still perform as well in the thermal spectrum. Thermal spectrum analysis may also have the potential to detect a backpack underneath a garment such as a rain poncho, provided the garment is thin enough to not dissipate heat being radiated through it by too great a degree.

A. Methods

A.1 Haar Face Detector

To robustly detect the players face haar-like feature detection was used. The haar classifier works by looking for rectangles with distinctive patterns as indicated in figure 1. In a human face certain regions are darker than others, the eyes for instance are much darker than the check region. Hence rectangle (a) laid over an eye check paid in a face would be a better fit than (b) or any of the other rectangles. Two, three and four region Haar classifiers have been defined.

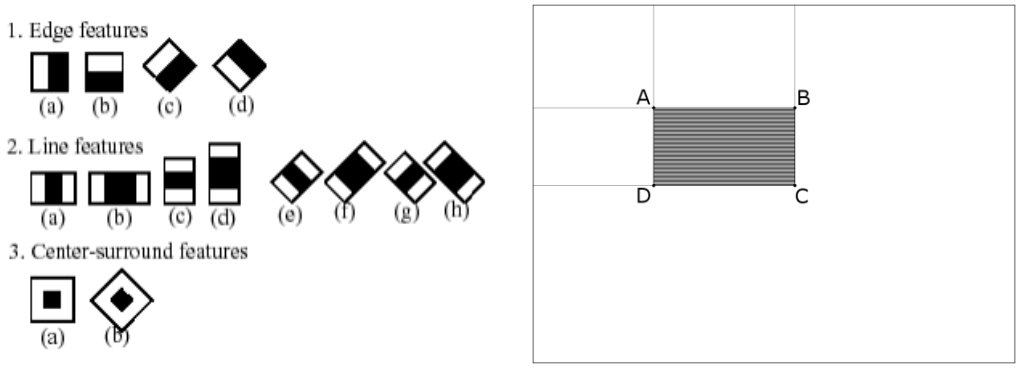


Figure A1 Haar Features left and the how to computer the intensity of a region using integral images on the right [1].

To work out if a certain Haar-like feature is present within the image integral images were computed. This consists of a matrix where each value is the sum of the intensity of all the pixels above and to the left of the current pixel. Hence to work out the intensity of a certain rectangular region within the image we only need to access and store four values as indicated below.

$$Region\ Intesntiy = I(C) + I(A) - I(B) - I(D)$$

This greatly increases the computational speed of working out the intensity of a region. A two region haar classifier only needs six look ups, a three region needs eight look ups and a four region classifier only needs nine lookups.

Work has been done on using additional Haar-classifiers tilted at 45^{o} such as indicated by (c) and (d) in the edge features in fig. 1 by Lienhart and Maydt. This was found to reduce the false positive rate in the region of 10%-12.5% at the expense of computing a second integral image as indicated in figure 2. This integral image is identical to the first except everything now rotated by 45 ^{o}. This integral image is used with two and three region haar-features but not four region ones.

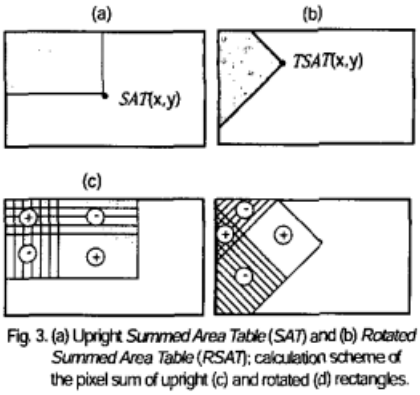


Fig. 3. (a) Upright *Summed Area Table* (*SAT*) and (b) *Rotated Summed Area Table* (*RSAT*); calculation scheme of the pixel sum of upright (c) and rotated (d) rectangles.

Figure A2 Tilted 45 degree Haar Calculation [37]

Further work was done one enabling haar-features to be calculated at any angle by Messom and Barczak however it was found to be computationally expensive. In addition detection algorithms then to use low resolution images resulting in rounding errors preventing the use of rotated haar classifiers.

This is known as a weak classifier as it will not always be correct as rectangles laid over a coloured image will not directly correspond to a perfect black and white rectangle. Hence we only get a confidence value for how well each rectangle fits the image. By setting a threshold and accepting confidence levels over a certain value we are quite likely to get in-correct classifications. These inaccuracies can be heightened by changing levels of light and interfering background colours. In addition some of these patterns are likely to occur naturally in the background.

To make the classifier more robust we turn it into a strong classifier using a method known as adaboost. This checks for multiple features and only returns a classification if all of them agree on the object. This reduces the number of in-correct classifications however to compute all of the haar features for a 24x24 subpixel region would require examining upwards of 30,000 features.

Hence we combine all of the haar features into a cascade lf several stages. For each region the computationally light haar classifiers are checked first. Following this more computational intensive features are checked for only on the regions that passed the initial test. By combining into several stages we can keep the computation down to a minimum for an image with an haar feature returning false eliminating that region from more checks.

A.2 Colour Spaces

Unlike gradient based methods which were analysed using only the luminance channel of an image K-Means takes place on a colour image. Hence there are several different colour spaces that can be used for clustering. As clustering can be applied to one, two or more channels various combinations of different colour spaces and different channels were tried to try and find the best cluster classification.

- BGR – This is the standard colour space and uses three channels to represent the intensity of the blue, green and red components of an image respectively.

- XYZ – This three channel colour space is designed to represent pixel values in a way the human eye better understands them. The Y channel is the luminance of the image, Z is approximate to the blue stimulation perceived by the human eye and X is a mix of the remaining values chosen to be non-negative.
- YCbCr – This three channel colour space is represented by the luminance channel Y. Cb and Cr are the differences of the blue and red components within the image with respect to Y. As shadows had a tendency to cause two clusters as shown in *fig. 66* in BGR images, clustering was attempted using only the Cb and Cr channels of an image in YCbCr.
- Luv and Lab and similar to YCbCr but derived from the XYZ colour space instead of the BGR. Once again cluster was attempted using only the ab and uv channels.
- HSV and HLS – These are two colour spaces that attempt to represent information about a pixel with respect to colour and luminance more intuitively. Hue is a circular representation of the colour space values of a pixel. Saturation is the intensity of that colour, for instance is it a light or dark green. The third channel is Lightness or Value both of which represent luminance information, the difference is that lightness goes from black to white with the maximum value of colour in the midpoint while Value goes from black to the maximum value of colour. Clustering was attempted using only the hue channel of both of these image to see if straps could be differentiated by only their colour value.

After experimentation clustering only selected channels of an image was found to provide no benefit to the system over conventional RGB or XYZ clustering. Clustering using the relative colour information from YCbCr, Luv and Lab as shown *fig. 66* even failed to remove clusters aligning to shadows rather than actual colour changes in the image. Clustering using the Hue channel only was a complete failure as shown by the last two images

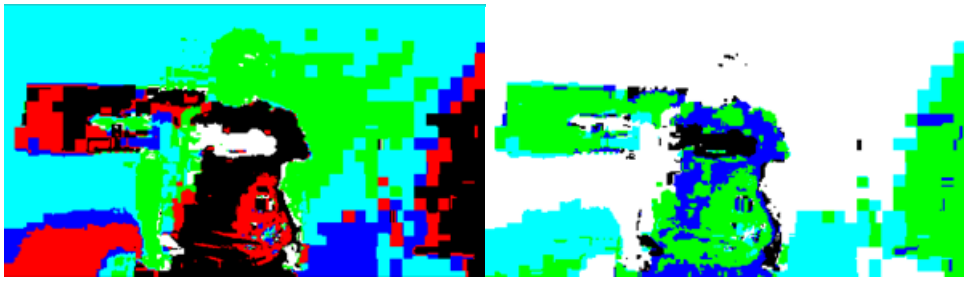


Figure 67: Clustering in different colour spaces, Clockwise from top left: Input, BGR, YCbCr, Luv, HSV, HLS, Lab, XYZ.

A.3 K-Means++

The standard K-Means algorithm requires cluster canters to be seeded. If the user has time to process the data they can provide custom seeds for the algorithm. However this is not beneficial for fully automated systems which are more likely to make use of random cluster seeding. The K-Means algorithm is vulnerable to bad initial seeding and will not always reach optimal clustering for an image. For instance take a rectangle which its major axis aligned to the x-axis and its minor aligned to the y-axis. If there are four points at each corner an optimal clustering would put the left two points in one cluster and the right two in the second. However if the initial cluster canters were at the endpoints of the minor axis K-Means would converge with the top two in one cluster and the bottom two points in a second cluster. In addition bad clusters will increase the execution time of the algorithm as it takes more iterations to correct the bad clusters.

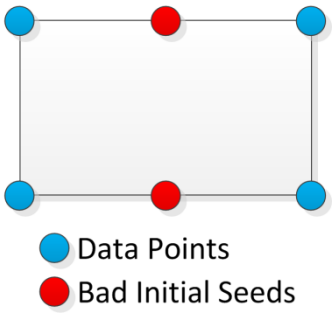


Figure 68: Bad initial seeds for K-Means clustering

To solve this problem [35] develop an initialisation algorithm to choose better initial clusters. The algorithm works of the principle that spreading out clusters results in good initial seeds and operates as follows:

1. The initial cluster centre is chosen randomly from all data points.
2. The next cluster is chosen from the remaining point with probability proportional to the square distance of each point to the nearest cluster centre.
3. Step 2 is repeated until all clusters are assigned.

According to the literature this significantly reduces the likelihood of K-Means producing bad centres

A.4 Detailed Annotation System

The first system stored for each frame the number of backpacks present and for each backpack the following information:

- bool: Is left strap visible
- bool: Is left strap occluded
- bool: Is right strap visible

- bool: Is right strap occluded

For every twentieth frame two points indicating the top left and bottom right of a bounding rectangle around the number of present straps for the backpack.

As unique test data was created for this project the annotations for this data also had to be created from scratch. This required the creation of a custom program that would play through each video frame by frame requiring the author to manually input all of the information for the required flags. This was a time consuming process and also extremely prone to operator error due to fatigue relating to time required to process all videos. It was also an unnecessary amount of information to be creating for each video.

A.5 Attached CD

Please Find attached the source code from the project as well as sample success and failure videos.

B. Results

B.1 Parameter Tuning Approach Four

		abs				Recall		Precision				
		TP	FP	FN	TN	TPR	TNR	PPV	NPV	Accuracy	F1	Matthews
A	*0.3*	4	1	18	21	0.182	0.955	0.800	0.538	0.568	0.296	0.215
	0.5	11	1	11	21	0.500	0.955	0.917	0.656	0.727	0.647	0.510
	0.6	10	3	12	19	0.455	0.864	0.769	0.613	0.659	0.571	0.349
	0.7	10	5	12	17	0.455	0.773	0.667	0.586	0.614	0.541	0.240
	0.8	6	5	16	17	0.273	0.773	0.545	0.515	0.523	0.364	0.052
B	*0.1*	6	5	16	17	0.273	0.773	0.545	0.515	0.523	0.364	0.052
	0.3	10	3	12	19	0.455	0.864	0.769	0.613	0.659	0.571	0.349
	0.4	10	4	12	18	0.455	0.818	0.714	0.600	0.636	0.556	0.293
	0.5	12	1	10	21	0.545	0.955	0.923	0.677	0.750	0.686	0.548
	0.8	7	1	15	21	0.318	0.955	0.875	0.583	0.636	0.467	0.354
C	*0.005*	13	3	9	19	0.591	0.864	0.813	0.679	0.727	0.684	0.472
	0.008	8	5	14	17	0.364	0.773	0.615	0.548	0.568	0.457	0.149
	0.01	11	4	11	18	0.500	0.818	0.733	0.621	0.659	0.595	0.336
	0.02	11	3	11	19	0.500	0.864	0.786	0.633	0.682	0.611	0.390
	0.03	9	5	13	17	0.409	0.773	0.643	0.567	0.591	0.500	0.195
D	*0.1*	9	5	13	17	0.409	0.773	0.643	0.567	0.591	0.500	0.195
	0.15	9	4	13	18	0.409	0.818	0.692	0.581	0.614	0.514	0.249
	0.2	11	3	11	19	0.500	0.864	0.786	0.633	0.682	0.611	0.390
	0.25	10	3	12	19	0.455	0.864	0.769	0.613	0.659	0.571	0.349
	0.3	11	3	11	19	0.500	0.864	0.786	0.633	0.682	0.611	0.390
E	*1*	1	1	21	21	0.045	0.955	0.500	0.500	0.500	0.083	0.000
	3	9	2	13	20	0.409	0.909	0.818	0.606	0.659	0.545	0.367
	4	7	2	15	20	0.318	0.909	0.778	0.571	0.614	0.452	0.282
	5	10	2	12	20	0.455	0.909	0.833	0.625	0.682	0.588	0.408
	8	10	4	12	18	0.455	0.818	0.714	0.600	0.636	0.556	0.293
F	*40*	7	3	15	19	0.318	0.864	0.700	0.559	0.591	0.438	0.217
	70	10	3	12	19	0.455	0.864	0.769	0.613	0.659	0.571	0.349
	80	10	2	12	20	0.455	0.909	0.833	0.625	0.682	0.588	0.408
	90	10	2	12	20	0.455	0.909	0.833	0.625	0.682	0.588	0.408
	120	8	3	14	19	0.364	0.864	0.727	0.576	0.614	0.485	0.262

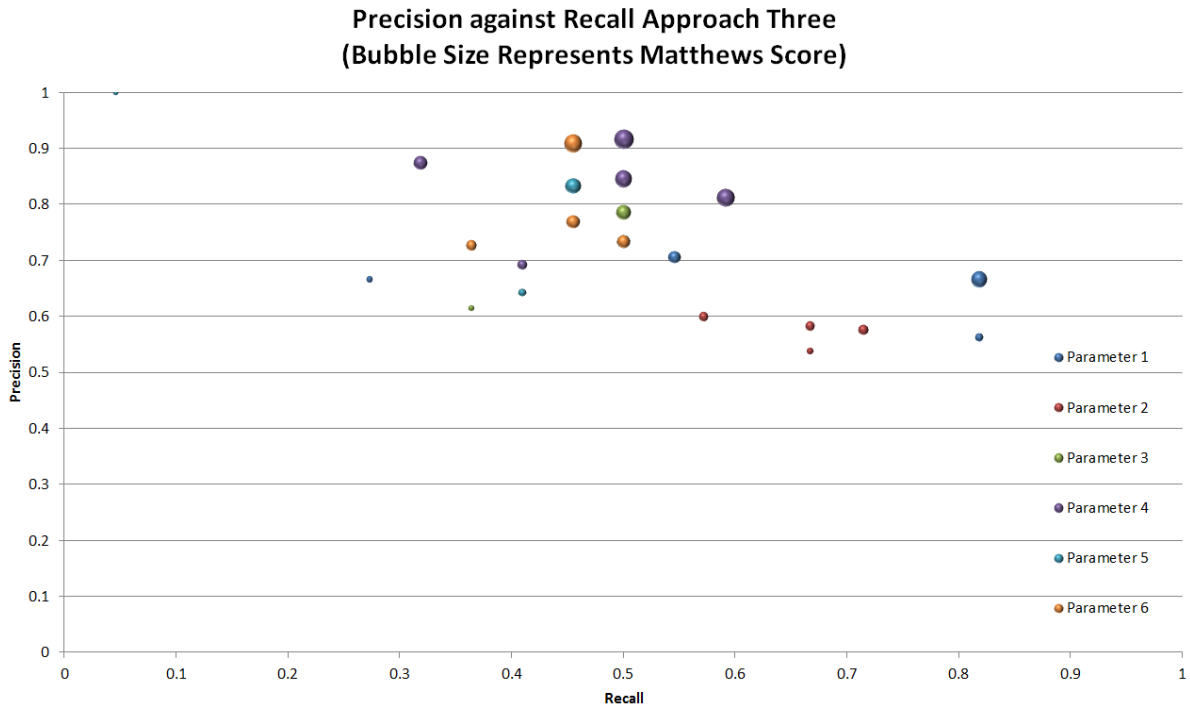
B.2 Parameter Tuning Approach Five

						Recall		Precision				
		TP	FP	FN	TN	TPR	TNR	PPV	NPV	Accuracy	F1	Matthews
A	*0.3*	18	14	4	8	0.818182	0.363636	0.5625	0.666667	0.590909	0.666667	0.20412415
	0.5	18	14	4	8	0.818182	0.363636	0.5625	0.666667	0.590909	0.666667	0.20412415
	0.6	18	9	4	13	0.818182	0.590909	0.666667	0.764706	0.704545	0.734694	0.42008403
	0.7	12	5	10	17	0.545455	0.772727	0.705882	0.62963	0.659091	0.615385	0.32673202
	0.8	6	3	16	19	0.272727	0.863636	0.666667	0.542857	0.568182	0.387097	0.16903085
B	*0.1*	12	8	9	16	0.571	0.667	0.600	0.640	0.622	0.585	0.23904572
	0.3	14	12	7	12	0.667	0.500	0.538	0.632	0.578	0.596	0.16834512
	0.4	15	11	6	13	0.714	0.542	0.577	0.684	0.622	0.638	0.25853001
	0.5	15	11	6	13	0.714	0.542	0.577	0.684	0.622	0.638	0.25853001
	0.8	14	10	7	14	0.667	0.583	0.583	0.667	0.622	0.622	0.25
C	*0.005*	13	3	9	19	0.591	0.864	0.813	0.679	0.727	0.684	0.47245559
	0.008	8	5	14	17	0.364	0.773	0.615	0.548	0.568	0.457	0.14944064
	0.01	11	4	11	18	0.500	0.818	0.733	0.621	0.659	0.595	0.33562431
	0.02	11	3	11	19	0.500	0.864	0.786	0.633	0.682	0.611	0.39036003
	0.03	9	5	13	17	0.409	0.773	0.643	0.567	0.591	0.500	0.19518001
D	*0.1*	11	1	11	21	0.500	0.955	0.917	0.656	0.727	0.647	0.51031036
	0.15	11	2	11	20	0.500	0.909	0.846	0.645	0.705	0.629	0.44832193
	0.2	13	3	9	19	0.591	0.864	0.813	0.679	0.727	0.684	0.47245559
	0.25	7	1	15	21	0.318	0.955	0.875	0.583	0.636	0.467	0.35355339
	0.3	9	4	13	18	0.409	0.818	0.692	0.581	0.614	0.514	0.24906774
E	*1*	1	0	21	22	0.045	1.000	1.000	0.512	0.523	0.087	0.15249857
	3	10	2	12	20	0.455	0.909	0.833	0.625	0.682	0.588	0.40824829
	4	10	2	12	20	0.455	0.909	0.833	0.625	0.682	0.588	0.40824829
	5	9	5	13	17	0.409	0.773	0.643	0.567	0.591	0.500	0.19518001
	8	11	4	11	18	0.500	0.818	0.733	0.621	0.659	0.595	0.33562431
F	*40*	10	3	12	19	0.455	0.864	0.769	0.613	0.659	0.571	0.34869484
	70	8	3	14	19	0.364	0.864	0.727	0.576	0.614	0.485	0.26243194
	80	11	4	11	18	0.500	0.818	0.733	0.621	0.659	0.595	0.33562431
	90	8	3	14	19	0.364	0.864	0.727	0.576	0.614	0.485	0.26243194
	120	10	1	12	21	0.455	0.955	0.909	0.636	0.705	0.606	0.47237749

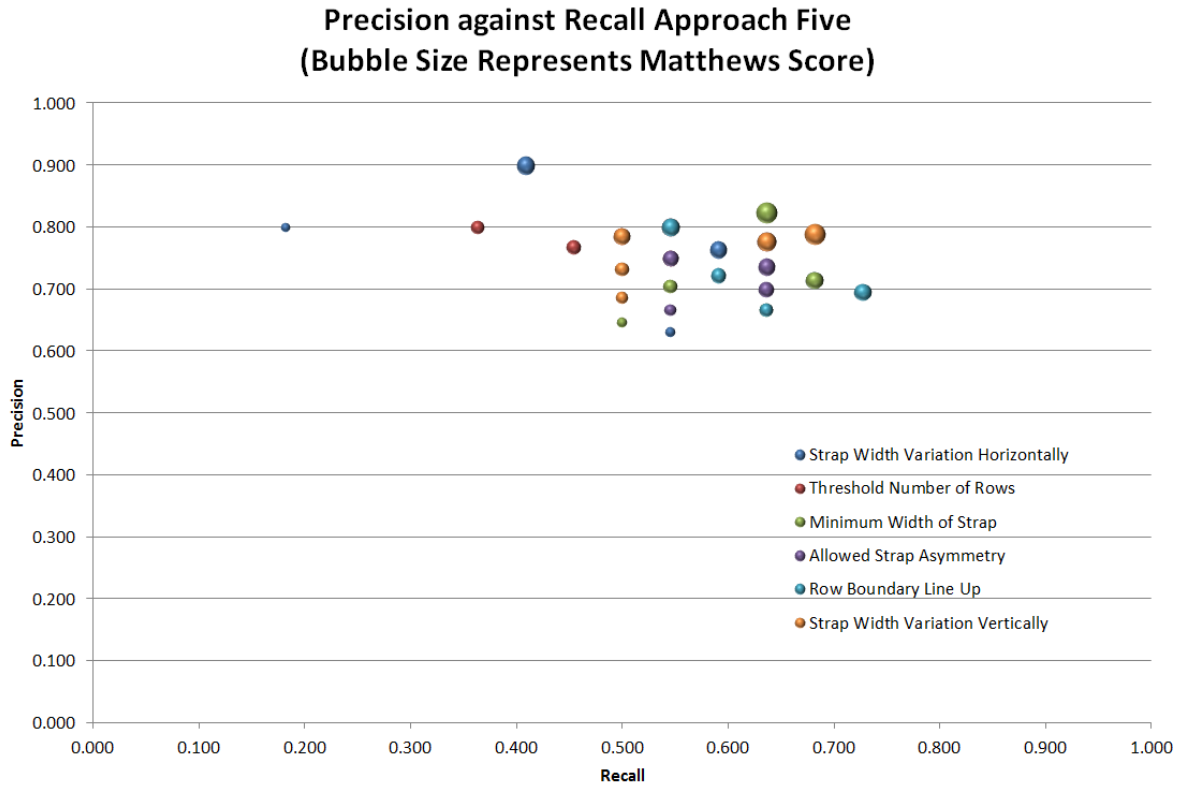
B.3 Best Results for Each Approach

Approach		TP	FP	FN	TN	Recall TPR	TNR	Precision PPV	NPV	Matthews	Accuracy	F1	Matthews
1	**1.25**	8	9	14	13	0.364	0.591	0.471	0.481	0.047	0.477	0.410	0.047
	1.40	7	6	15	16	0.318	0.727	0.538	0.516	0.050	0.523	0.400	0.050
	1.50	7	6	15	16	0.318	0.727	0.538	0.516	0.050	0.523	0.400	0.050
H/W	**1.60**	6	9	16	13	0.273	0.591	0.400	0.448	0.144	0.432	0.324	0.144
Ratio	**1.75**	6	7	16	15	0.273	0.682	0.462	0.484	0.050	0.477	0.343	0.050
2	**1**	10	4	12	18	0.455	0.818	0.714	0.600	0.293	0.636	0.556	0.293
	2	10	8	12	14	0.455	0.636	0.556	0.538	0.092	0.545	0.500	0.092
	3	13	9	9	13	0.591	0.591	0.591	0.591	0.182	0.591	0.591	0.182
Discarded	**4**	16	10	6	12	0.727	0.545	0.615	0.667	0.277	0.636	0.667	0.277
Lines	**5**	16	12	6	10	0.727	0.455	0.571	0.625	0.189	0.591	0.640	0.189
3	**0.500**	19	18	3	4	0.864	0.182	0.514	0.571	0.062	0.523	0.644	0.062
	0.400	19	17	3	5	0.864	0.227	0.528	0.625	0.118	0.545	0.655	0.118
	0.300	18	16	4	6	0.818	0.273	0.529	0.600	0.108	0.545	0.643	0.108
	0.200	18	12	4	10	0.818	0.455	0.600	0.714	0.293	0.636	0.692	0.293
Distance	**0.150**	16	10	6	12	0.727	0.545	0.615	0.667	0.277	0.636	0.667	0.277
between	**0.146**	17	10	5	12	0.773	0.545	0.630	0.706	0.327	0.636	0.667	0.327
two	**0.143**	17	10	5	12	0.773	0.545	0.630	0.706	0.327	0.636	0.667	0.327
Edges	**0.140**	17	9	5	13	0.773	0.591	0.654	0.722	0.370	0.636	0.667	0.370
	0.137	16	9	6	13	0.727	0.591	0.640	0.684	0.321	0.636	0.667	0.321
	0.134	16	9	6	13	0.727	0.591	0.640	0.684	0.321	0.636	0.667	0.321
	0.130	17	10	5	12	0.773	0.545	0.630	0.706	0.327	0.659	0.694	0.327
	0.120	16	10	6	12	0.727	0.545	0.615	0.667	0.277	0.636	0.667	0.277
	0.110	14	8	8	14	0.636	0.636	0.636	0.636	0.273	0.636	0.636	0.273
	0.100	12	5	10	17	0.545	0.773	0.706	0.630	0.327	0.659	0.615	0.327
	0.050	5	3	17	19	0.227	0.864	0.625	0.528	0.118	0.545	0.333	0.118
4	**0.500**	18	14	4	8	0.818	0.364	0.563	0.667	0.204	0.591	0.667	0.204
	0.400	18	14	4	8	0.818	0.364	0.563	0.667	0.204	0.591	0.667	0.204
	0.360	19	12	3	10	0.864	0.455	0.613	0.769	0.349	0.659	0.717	0.349
	0.300	18	9	4	13	0.818	0.591	0.667	0.765	0.420	0.705	0.735	0.420
Distance	**0.296**	18	9	4	13	0.818	0.591	0.667	0.765	0.420	0.705	0.735	0.420
between	**0.293**	18	9	4	13	0.818	0.591	0.667	0.765	0.420	0.705	0.735	0.420
two	**0.280**	18	10	4	12	0.818	0.545	0.643	0.750	0.378	0.705	0.735	0.378
Edges	**0.277**	18	9	4	13	0.818	0.591	0.667	0.765	0.420	0.705	0.735	0.420
	0.274	18	9	4	13	0.818	0.591	0.667	0.765	0.420	0.705	0.735	0.420
	0.270	18	9	4	13	0.818	0.591	0.667	0.765	0.420	0.705	0.735	0.420
	0.200	12	5	10	17	0.545	0.773	0.706	0.630	0.327	0.659	0.615	0.327
	0.100	6	3	16	19	0.273	0.864	0.667	0.543	0.169	0.568	0.387	0.169
5	**0.003**	11	3	11	19	0.500	0.864	0.786	0.633	0.390	0.727	0.647	0.390
	0.004	13	6	9	16	0.591	0.727	0.684	0.640	0.321	0.705	0.629	0.321
	0.005	15	3	7	19	0.682	0.864	0.833	0.731	0.555	0.727	0.684	0.596
Minimum	**0.006**	13	4	9	18	0.591	0.818	0.765	0.667	0.420	0.636	0.467	0.420
Width	**0.007**	13	6	9	16	0.591	0.727	0.684	0.640	0.321	0.614	0.514	0.321
6	**0.35**	33	18	2	4	0.943	0.182	0.478	0.800	0.198	0.000	0.614	0.198
	0.36	28	12	7	10	0.800	0.455	0.538	0.741	0.271	0.000	0.614	0.271
	0.37	28	8	7	14	0.800	0.636	0.636	0.800	0.440	0.860	0.674	0.440
Minimum	**0.38**	26	9	9	13	0.743	0.591	0.591	0.743	0.334	0.877	0.659	0.334
Width	**0.39**	25	8	10	14	0.714	0.636	0.610	0.737	0.346	0.860	0.559	0.346

B.4 PR Curve Approach Four



B.5 P-R Curve Approach Five



C. References

[1] K. Dawson-Howe, "Lectures Notes for CS4053 Vision," 2012.

[2] M. Valera and S. A. Velastin, "Intelligent distributed surveillance systems: a review," *Vision, Image and Signal Processing, IEE Proceedings -,* vol. 152, pp. 192-204, 2005.

[3] BSIA, "One surveillance camera for every 11 people in Britain, says CCTV survey," *The Telegraph,* 2013.

[4] M. Stroud, "In Boston bombing, flood of digital evidence is a blessing and a curse," *CNN International* 2013.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, pp. 886-893 vol. 1.

[6] I. Haritaoglu, R. Cutler, D. Harwood, and L. S. Davis, "Backpack: Detection of people carrying objects using silhouettes," *Computer Vision and Image Understanding,* vol. 81, pp. 385-397, 2001.

[7] G. Frankel, "London Hit Again with Explosions," *The Washington Post,* 2005.

[8] OpenCV, "Face Detection using Haar Cascades," *docs.opencv.org,* 2014.

[9] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, 2001, pp. I-511-I-518 vol.1.

[10] A. L. Kuranov, R. Pisarevsky, V., "An Empirical Analysis of Boosting Algorithms for Rapid Objects With an Extended SEt of Haar-like Features," *Intel Technical Report,* 2002.

[11] H. Dong-Chen and W. Li, "Texture Unit, Texture Spectrum, And Texture Analysis," *Geoscience and Remote Sensing, IEEE Transactions on,* vol. 28, pp. 509-512, 1990.

[12] P. KaewTraKulPong and R. Bowden, "An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection," in *Video-Based Surveillance Systems*, P. Remagnino, G. Jones, N. Paragios, and C. Regazzoni, Eds., ed: Springer US, 2002, pp. 135-144.

[13] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2004, pp. 28-31 Vol.2.

[14] A. B. Godbehere, A. Matsukawa, and K. Goldberg, "Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation," in *American Control Conference (ACC), 2012*, 2012, pp. 4305-4312.

[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, pp. 1627-1645, 2010.

[16] H. Cho, P. E. Rybski, and W. Zhang, "Vision-based bicycle detection and tracking using a deformable part model and an EKF algorithm," in *13th International IEEE Conference on Intelligent Transportation Systems, ITSC 2010, September 19, 2010 - September 22, 2010*, Funchal, Portugal, 2010, pp. 1875-1880.

[17] H. Cho, P. E. Rybski, and W. Zhang, "Vision-based 3D bicycle tracking using deformable part model and interacting multiple model filter," in *2011 IEEE International Conference on Robotics and Automation, ICRA 2011, May 9, 2011 - May 13, 2011*, Shanghai, China, 2011, pp. 4391-4398.

[18] K. Takahashi, Y. Kuriya, and T. Morie, "Bicycle detection using pedaling movement by spatiotemporal Gabor filtering," in *2010 IEEE Region 10 Conference, TENCON 2010, November 21, 2010 - November 24, 2010*, Fukuoka, Japan, 2010, pp. 918-922.

[19] I. Haritaoglu, R. Cutler, D. Harwood, and L. S. Davis, "Backpack: detection of people carrying objects using silhouettes," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, 1999, pp. 102-107 vol.1.

[20] I. Haritaoglu, D. Harwood, and L. S. Davis, "W⁴: Who? When? Where? What? A real time system for detecting and tracking people," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, 1998, pp. 222-227.

[21] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 22, pp. 781-796, 2000.

[22] B. DeCann and A. Ross, "Gait curves for human recognition, backpack detection, and silhouette correction in a nighttime environment," in *Biometric Technology for Human Identification VII, April 5, 2010 - April 6, 2010*, Orlando, FL, United states, 2010, pp. The Society of Photo-Optical Instrumentation Engineers (SPIE).

[23] CASIA, "Chinese Academy of Sciences Gait Databse, Dataset C," 2005.

[24] L. D. Chiraz BenAbdelkader, "Detection of People Carrying Objects: a Motion-based Recognition Approach," 2002.

[25] D. Tao, X. Li, X. Wu, and S. J. Maybank, "Human carrying status in visual surveillance," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006, June 17, 2006 - June 22, 2006*, New York, NY, United states, 2006, pp. 1670-1677.

[26] D. Damen and D. Hogg, "Detecting carried objects in short video sequences," in *10th European Conference on Computer Vision, ECCV 2008, October 12, 2008 - October 18, 2008*, Marseille, France, 2008, pp. 154-167.

[27] D. Damen and D. Hogg, "Detecting Carried Objects from Sequences of Walking Pedestrians," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 34, pp. 1056-1067, 2012.

[28] C. Chi-Hung, H. Jun-Wei, T. Luo-Wei, C. Sin-Yu, and F. Kuo-Chin, "Carried object detection using ratio histogram and its application to suspicious event analysis," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 19, pp. 911-916, 2009.

[29] A. Branca, M. Leo, G. Attolico, and A. Distante, "Detection of objects carried by people," in *International Conference on Image Processing (ICIP'02), September 22, 2002 - September 25, 2002*, Rochester, NY, United states, 2002, pp. III/317-III/320.

[30] C. Teck Wee, K. Leman, W. Hee Lin, P. Nam Trung, R. Chang, N. Dinh Duy*, et al.*, "Sling bag and backpack detection for human appearance semantic in vision system," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, 2013, pp. 2130-2135.

[31] D. Corrigan, "Image and Video Quality Assessment Subjective Testing," *5C2 Digital Media,* 2014.

[32] B. S. Baldi P., Chauvin Y., Andersen C. A., Nielsen H., "Assessing the Accuracy of Prediction Algorithms for Classification: an Overview," *Bioinformatics,* 2000.

[33] R. Eshel and Y. Moses, "Homography based multiple camera detection and tracking of people in a dense crowd," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1-8.

[34] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics),* vol. 28, pp. 100-108, 1979.

[35] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," presented at the Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, New Orleans, Louisiana, 2007.

[36] O. S. M. R. Lopez, V. Tyrsa, "Machine Vision: Approaches and Limitations," 2008.

[37] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, 2002, pp. I-900-I-903 vol.1.