University of Dublin, Trinity College

Tracking with Height Estimation in a Autonomous Surveillance System

A dissertation submitted to the University of Dublin, in part fulfillment of the requirements for the degree of Master of Computer Science.

Author:
Sinead Plunkett

Supervisor:
Dr. Kenneth Dawson-Howe

Submitted to the University of Dublin, Trinity College, May, 2015.

Declaration

I, Sinead Plunkett, 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:

Summary

The goal of this project is to correctly estimate the height of people on video to be used to aid in people tracking and identification.

The approach of this dissertation is to use an object of a known size to estimate height of standing people in a scene to better track and re-identify them as they move between fields of view. This would help solve the occlusion problem when one person walks in front of another. This approach would be able to calculate the size of any detected object in the scene by comparing it to the size of the known object which would normally be found in the scene.

In systems where there is a large number of cameras, it is difficult and time-consuming to calibrate the cameras. Accordingly, automatic camera calibration must be a key technology for multiple camera systems to be practical. Because this object would be seen normally in the scene it would not be necessary to manually calibrate the cameras.

A common object seen in outdoor scenes, especially around a college campus, is a bicycle wheel (attached to a bicycle). Most road and racing bicycles today use 622 mm diameter rims [Brown and Allen, 2014]. By first identifying the bicycle wheel and its position in the two dimensional representation of the scene, it would be possible to calculate the real world height or size of any other identified object (in this case a person) at that point in the image.

The system first finds all moving objects in the video. Each moving object is identified as either a bicycle, person or unidentified object (neither bicycle nor person). If the moving object was found to be a bicycle, its wheels are found and measured in pixels. The pixel height of the wheels are stored as well as each wheel's position. If the moving object was found to be a person, that person's real world height is calculated by first checking if the person is standing on a

point where a wheel has been found, if it is then the real world height of the person can be calculated.

For this a background subtraction algorithm was used to detect moving objects, as well as a people detector. A bicycle wheel detector was also needed.

A system was developed to test this using C++ and OpenCV. The developed system depends on lighting not changing in the video.

Acknowledgement

I would like to thank my supervisor, Dr. Kenneth Dawson-Howe, for his help and guidance during this project.

I want to thank my family, especially my parents for their support and encouragement.

Special thanks to: Aine Bannon, Katharine Burton, Ellen Burke, Damian Griffith-Bourke, Karen Conboy, Jamie Flynn and Luke Marjoram for proof reading, double checking my calculations, helping with testing and bringing me tea.

Contents

Associated code from project is on attached CD

1 Introduction

The goal of this project is to correctly estimate the height of people on video to be used to aid in people tracking.

For this dissertation the approach is to use an object of a known size to estimate height of standing people in a scene to better track and re-identify them as they move between fields of view. This would help solve the occlusion problem when one person walks in front of another. If both their sizes were known beforehand, when they move away from each other it would be easy to see who is who. This approach would be able to calculate the size of any detected object in the scene by comparing it to the size of the known object which would normally be found in the scene.

In systems where there is a large number of cameras, it is difficult and time-consuming to calibrate all of them. Accordingly, automatic camera calibration must be a key technology for multiple camera systems to be practical.

Because this object would be found normally in the scene it would not be necessary to manually calibrate the cameras.

A common object found in outdoor scenes, especially around a college campus, is a bicycle wheel (attached to a bicycle). Most road and racing bicycles today use 622 mm diameter rims [Brown and Allen, 2014]. By first identifying the bicycle wheel and its position in the two dimensional representation of the scene, it would be possible to calculate the real world height or size of any other identified object (in this case a person) at that point in the image.

This is as simple as dividing the real world diameter (622 mm) by the height of the detected wheel in pixels and multiplying the result by the height of the detected person in pixels.

As this uses a video at least one of these objects will move over time. While moving its size should not change in the real world but it would appear to change in the images as the object moves towards or away from the camera. If another object were to cross paths with this moving object, this second object's size could be calculated as so on.

Drawbacks to this approach include needing a bicycle wheel in the scene at some point.

A further issue is that the occlusion problem is not solved if, before occlusion, the real world height of the people has not been calculated, or if after occlusion, it cannot be calculated.

1.1 Motivation

In this section I will discuss the motivation for attempting to estimate standing people's real world height in a video surveillance network.

1.1.1 Video Surveillance

Surveillance is used by governments and law enforcement to maintain social control, recognize and monitor threats, and prevent or investigate criminal activity. It also has a number of other applications such as security in the home, workplace, casinos, banks etc. Video surveillance has played a crucial role in tracing the movements of suspects or victims and is widely regarded by anti-terrorist officers as a fundamental tool in tracking terrorist suspects [Yesil, 2006]. In addition, surveillance is increasingly used in retail and business process analysis, and marketing.

As surveillance cameras are becoming increasingly inexpensive, their numbers are increasing. There is an estimated 100 million cameras in operation worldwide [Vos, 2015]. The manpower required to monitor and analyse these cameras is costly, however. Consequently the videos from these cameras are usually unmonitored or, at best, rarely.

In addition, most security and surveillance systems are made up of more than one camera. Use of several cameras can provide a more complete history of a person's actions in an environment, as it is not possible for one camera to provide sufficient coverage of the environment because of its limited field of view.

Multiple camera based visual surveillance systems can be extremely helpful because the surveillance area is expanded and multiple view information can help overcome occlusion. Tracking with a single camera easily generates ambiguity due to occlusion or depth. This ambiguity may be eliminated from another view. However, visual surveillance using multiple cameras also brings problems such as camera installation (how to cover the entire scene with the minimum number of cameras), camera calibration, object matching, automated camera switching, and data fusion [Hu et al., 2004]. The video process of surveillance systems has inherited difficult challenges while using a computer vision application, i.e., illumination variation, viewpoint variation, scale (view distance) variation, and orientation variation [Hu et al., 2004].

As the costs associated with video hardware are decreasing, the number available streams of a video surveillance system is easily increased. The increased number of streams, while allowing a more thorough coverage of an area, further increases the amount of footage that must be sifted through by humans trying to recognize important events or incidents and track people through the monitored space.

Cameras provide the ability to review events at a later date as well as real time. This recording is important as it can be reviewed after an incident occurs. However, a huge volume of footage is likely to be produced and after an incident occurs all of this footage must be reviewed.

Finding available human resources to review the footage is expensive and is prone to error due to fatigue or negligence. The dependence on these human resources prevents surveillance systems from realizing their full potential. Therefore, the development of accurate and efficient automatic video analysis systems for monitoring human activity is important.

Current surveillance systems are unintelligent and unable to recognize what they are seeing. An ideal solution to this problem would be a system that can recognize and respond to what it sees without requiring an operator.

Surveillance cameras could be a far more useful tool if instead of passively recording footage, they can be used to detect events that require attention in real time, and take action as needed. This is the goal of automated visual surveillance: to obtain a description of what is happening in a monitored area, and then to take appropriate action based on that interpretation.

Visual processing of people includes detection, tracking, recognition, and behaviour interpretation. It is a key component of automatic video analysis systems and this computer analysis of human actions is gaining increasing interests, especially in video surveillance.

1.1.2 People Tracking

Object and People tracking has received a large amount of attention due to its numerous potential applications such as video surveillance, human activity analysis, traffic monitoring etc.

A common surveillance activity is to track important people, or people exhibiting suspicious behaviour, as they move from camera to camera.

Tracking multiple people from standard cameras is challenging. Humans can greatly change their appearance according to posture, clothing and lighting conditions. Targets can have similar appearances, such as a pedestrian in a crowded scene. Object tracking for targets with distinctive appearance is comparatively easier, but defining features that describe people moving in large scenarios is a complex task.

Occlusions (obstructions of a view) also present a problem where several people are involved as the tracker must separate the targets and assign them correct labels. Most security and surveillance systems are made up of more than one camera. There are two reasons for this. First, it is not possible for one camera to provide sufficient coverage of the environment because of its limited field of view. Second, the occlusion problem can be tackled using multiple cameras, observing the scene with different points of views. Multiple cameras can provide a more complete history of a person's actions in an environment. To take advantage of additional cameras, it is necessary to establish correspondence between different views.

Algorithms that can only handle the stream of one camera cannot cope in scenarios of a large group of people or heavy traffic.

The problem of who enters the area and/or engages in an abnormal or suspicious act under surveillance is of increasing importance for visual surveillance. Human face and gait are now regarded as the main biometric features that can be used for personal identification in visual surveillance systems.

1.2 Project Goals

The goal of this project is to design and implement a prototype people tracking system that can correctly estimate the height of people on video. This would have various uses:

1. It would be useful in crime prevention and investigation to know a suspects height to help identify him/her from the video footage.

2. In a multiple camera surveillance network, this would help to re-identify people (where cameras overlap) from different camera angles at the same person and (where cameras do not overlap) re-identify people as they move from one area to the next.

3. Occlusion in Computer Vision is the obstruction of the view, for example if a person was to walk behind another person, the first person would be hidden from the camera. The problem with occlusion is, in this example, when both persons walk apart the system is unable to determine who is

who. To be able to calculate a moving object's real world height would help to help solve the occlusion problem, if their real world heights were calculated both before and after occlusion, these heights could be matched so the tracking system could assign them correct labels.

1.3 Project Summary

The goal of this project is to correctly estimate the height of people on video to be used to aid in people tracking and identification.

The approach of this dissertation is to use an object of a known size to estimate height of standing people in a scene to better track and re-identify them as they move between fields of view. This would help solve the occlusion problem when one person walks in front of another. This approach would be able to calculate the size of any detected object in the scene by comparing it to the size of the known object which would normally be seen in the scene.

In systems where there is a large number of cameras, it is difficult and time-consuming to calibrate the cameras. Accordingly, automatic camera calibration must be a key technology for multiple camera systems to be practical. Because this object would be seen normally in the scene it would not be necessary to manually calibrate the cameras.

A common object seen in outdoor scenes, especially around a college campus, is a bicycle wheel (attached to a bicycle). Most road and racing bicycles today use 622 mm diameter rims [Brown and Allen, 2014]. By first identifying the bicycle wheel and its position in the two dimensional representation of the scene, it would be possible to calculate the real world height or size of any other identified object (in this case a person) at that point in the image.

The system first finds all moving objects in the video. Each moving object is identified as either a bicycle, person or unidentified object (neither bicycle nor person). If the moving object was found to be a bicycle its wheels are found and measured in pixels. The pixel height of the wheels are stored as well as each wheel's position. If the moving object was found to be a person, that person's real world height is calculated by first checking if the person is standing on a point where a wheel has been found, if it is then the real world height of the person can be calculated.

For this a background subtraction algorithm will be used to detect moving objects, as well as a people detector. A bicycle wheel detector will also be needed.

1.4 Overview of this Report

In Section 1 (Introduction) the background, motivation, goals and a brief summary of this dissertation is given. The problems that it hopes to solve are also discussed.

Section 2 (Literary Review) gives a brief description of other research that has been done in relation to this project. This focuses on both camera networks and biometrics.

Section 3 (Design) discusses the idea which this project aims solve the problems outlined in Section 1.

The implementation of the system used to test this idea is explained in Section 4 (Implementation). This gives a detailed description of the developed application.

Results are given in Section 5 (Results), along with an evaluation of those results and, finally, Section 6 (Conclusion) explores the future work and conclusion of the dissertation.

2 Literary Review

2.1 Introduction

This chapter will briefly describe other research that has been completed that is relevant to this dissertation.

Many different approaches to the multiple people tracking problem exist, yet there are still few satisfactory methods of solving the multi-target occlusion problem. This section aims to give examples of the research completed which has attempted to solve this problem.

A large amount of work has been done which considers overlapping or partially overlapping cameras as well as disjoint cameras. Biometrics is a very active area of research, there is much research to do with this.

2.2 Camera Networks

[Khan and Shah, 2003] present a method which aims to solve the consistent-labelling across different cameras problem, for cameras with overlapping FOVs (Field of View). It does not require camera calibration. They first discover the FOV lines, and use a labelling scheme based on the smallest distance between a line and the middle point of the bottom of the boundary box of an object. They show that when FOV lines are recovered, the consistent-labelling problem can be solved successfully. But it is also stated that the feet of the people, and the portions of the ground have to be visible, and there needs to be enough traffic across a particular FOV line to be able to recover it.

[Topcu et al., 2014] use epipolar geometry which takes advantage of the different views from the multiple cameras to form one to solve the occlusion problem. They also make use of a particle filter tracker which is initiated upon detection of a new person and terminated when they leave the scene.

2.3 Biometrics

Human identification biometrics systems may originate from real-life criminal and forensic applications. Some methods, such as fingerprinting and face recognition, already proved to be very efficient in computer vision based human recognition systems.

Biometrics is the science and technology of measuring and analysing biological data to identify a person. Examples include DNA, shape of the ear, faces, fingerprints, hand geometry, irises, pattern of keystrokes on a keyboard, signature and speech.

According to [Phillips et al., 1998] in order for a biometric to be effective it should have the following four properties:

1. **Universality** - All members of population being identified should possess the biometric.

2. **Uniqueness** - Biometric signature should be different for all members of the population.

3. **Invariance** - The signature should be invariant under the conditions that it will be collected.

4. **Resistance** - Biometrics should be resistant to potential countermeasures.

Some biometrics can be collected without a persons knowledge, for example speech, signature, keystrokes and ears. Others, for example, fingerprints and hand geometry, require a persons cooperation. This distinction is important. Potential applications for non-intrusive biometrics include remote surveillance and monitoring border crossings and airport security while potential applications for cooperative biometrics include access control and authentication.

Faces, fingerprints, irises and ears require image processing pattern recognition and computer vision techniques to be implemented. Whereas fingerprints, keystrokes, signature and speech fall into the domains of signal processing and pattern recognition. Recent reports have looked at combining multiple biometrics, audio/video, faces, fingerprints etc.

2.3.0.1 Fingerprints To date fingerprints have been primarily used by law enforcement applications and for background clearances with the availability of an inexpensive live scanner. Fingerprints are applied to some verification applications.

Images captured by a live scan reader can be directly inputted into an AFIS (Automated Fingerprint Identification System) for subsequent processing. Benefits of this new technology include the elimination of ink, determination of image quality before recording multiple copies of the same image from a single scanning and the immediate creation of a file containing the electronic fingerprint image [Phillips et al., 1998].

Cooperation is needed for the collection of this biometric. It is therefore not suitable for a non-intrusive video surveillance system.

2.3.0.2 Palm Prints Palm print recognition is another emerging biometrics technique also primarily used by law enforcement applications [Chorás, 2007]. A palm print is a good biometric identifier because of its universality, uniqueness and invariance. Human palm prints are rich in texture and are stable and show high accuracy in representing the identity of each individual [Chen et al., 2001]. Thus, they have been commonly used in law enforcement and forensic environments.

Again, cooperation and knowledge is needed for the collection of this biometric. It is therefore not suitable for a non-intrusive video surveillance system.

2.3.0.3 Face A fully automatic system would detect and identify a face in an image or video sequence without human intervention.

Face recognition relies on the ability to handle variations in lighting and pose, and the ability to detect and track people in video. One of the most challenging problems in face recognition is to account for ageing which will be useful in recognizing missing and exploited children[Phillips et al., 1998].

Face recognition relies on high quality image capture (high resolution), so features can be distinguished. Storage of high quality images require more space and is expensive, making this an unattractive choice.

Eyes For irises the major research direction is acquiring the irises at increasingly greater distances [Phillips et al., 1998].

Lips Lip recognition has not been extensively researched so far, but some very promising results were achieved by [Gomez et al., 2002].

2.3.0.4 Ear Human ears have been used as major feature in forensic science (for example in airplane crashes)[Chorás, 2007]. Police and forensic specialists use ear prints as a standard proof of identity [Hoogstrate et al., 2001] [Kasprzak, 2003] [Pasescu and Tanislav, 1997]. There are many advantages of using the ear as a source of data for human identification. Firstly, the ear is one of the most stable human anatomical feature, as proven by [Iannarelli, 1989] (it does not change considerably during human life). The ears are our sensors for sound and therefore they are usually visible (not hidden underneath anything) to enable good hearing.

Again, the person must be fairly close to the camera and the image quality must be good enough (high resolution) to distinguish these features in order for this method to work.

2.3.0.5 Gait Because it can be measured "at a distance" when other biometrics are obscured, there is an increased interest in using gait features (the manner in which a person walks) for human identification in surveillance applications.

A unique advantage of gait as a biometric is that it offers potential for recognition at a distance or at low resolution, when other biometrics might not be perceivable [Nixon et al., 1999].

Human walking involves rhythmic up-and-down displacement of the upper body (from pelvis to head), hence the apparent bobbing of the head [Inman et al., 1981] [Rose et al., 2006]. Furthermore, these vertical movements must occur in a smooth sinusoidal manner for the conservation of energy [Inman et al., 1981]. This movement was found to differ between people. Its recognition capability is supported by studies in other areas such as medicine, mathematics and psychology, which continue to suggest that gait is unique [Nixon and Carter, 2004].

Another advantage to using gait recognition is that it is difficult to disguise gait without hampering progress, which is of particular interest in scene of crime analysis. Recognition can be based on the (static) human shape as well as on movement, suggesting a richer recognition cue. Furthermore, gait can be used when other biometrics are obscured criminal intent might motivate concealment of the face, but it is difficult to conceal and/or disguise motion as this generally impedes movement.

2.3.0.6 Height In [Kispál and Jeges, 2008], the height of a person with visible feet and head is estimated from an already calibrated camera. [Criminisi et al., 1999], [Hoiem et al., 2008], and [Lalonde et al., 2007] describe the measurement of various objects (including people) rooted on the ground plane. [Gallagher et al., 2009] uses a calibrated camera for measuring the height of people in the scene. Height is measured by jointly inferring across anthropometric dimensions (human body measurements in relation to each other), age, and gender using publicly available statistics. The height estimate provides context for recognizing the age and gender of the subject, and likewise age and gender conditions the distribution of the anthropometric features for estimating height.

2.4 Conclusion

This chapter discussed previous research in the area of people identification and tracking across a video surveillance network. These discoveries will aid in the work for this dissertation.

There is much more research published regarding this topic. Those mentioned above were read for this dissertation. It is clear to see that there is still work to be done to solve the people tracking problem.

3 Design

This chapter will discuss design decisions, with an overview and detailed look at the theory behind this new idea. Key challenges surrounding the project will be outlined in this chapter.

3.1 Introduction

The goal of this project is to correctly estimate the height of people on video to be used to aid in people tracking.

The approach of this dissertation is to use an object of a known size to estimate height of standing people in a scene to better track and re-identify them as they move between fields of view. This would help solve the occlusion problem when one person walks in front of another. If both their sizes were known beforehand, when they move away from each other it would be easy to see who is who. This approach would be able to calculate the size of any detected object in the scene by comparing it to the size of the known object which would normally be seen in the scene.

In systems where there is a large number of cameras, it is difficult and time-consuming to calibrate the cameras. Accordingly, automatic camera calibration must be a key technology for multiple camera systems to be practical.

Because this object would be seen normally in the scene it would not be necessary to manually calibrate the cameras.

A common object seen in outdoor scenes, especially around a college campus, is a bicycle wheel (attached to a bicycle). Most road and racing bicycles today use 622 mm diameter rims [Brown and Allen, 2014]. By first identifying the bicycle wheel and its position in the two dimensional representation of the scene, it would be possible to calculate the real world height or size of any other identified object (in this case a person) at that point in the image.

3.2 Overview

The system first finds all moving objects in the video. Each moving object is identified as either a bicycle, person or unidentified object (neither bicycle nor person). If the moving object was found to be a bicycle its wheels are found and measured in pixels. The pixel height of the wheels are stored as well as each wheel's position. If the moving object was found to be a person, that person's real world height is calculated by first checking if the person is standing on a point where a wheel has been found, if it is then the real world height of the person can be calculated.

To create this system a people detector is needed. For this a background subtraction algorithm will be used as well as the HOG (Histogram of Orientated Gradients, described in section 4.3.3) people detector. With background subtraction the difference is calculated between the current frame and a background image to find the moving objects in a scene. This requires the acquisition and maintenance of the background image. For this the Gaussian mixture model [Stauffer and Grimson, 1999] will be used, which is able to deal with objects which move slightly (for example leaves moving in the wind). The moving pixels will be joined to form a moving blob. These blobs will need to be identified as people, this will be done using the HOG [Dalal and Triggs, 2005] people detector.

A bicycle wheel detector will also be needed. For this I made my own.

To increase the area in which the person's height can be detected, an estimation between 2 points where wheels have been found can be calculated.

To cover a much wider area, a third bicycle wheel's height must be detected. The expected height can be calculated at any point within the triangle that is formed between the three known points. This can be done by finding the line that passes through one of the triangles vertices, the point where the height needs to be calculated and and of the edges of the triangle (where the height at the point or intersection can be estimated with the methods mentioned above).

3.3 Design Problems

The design assumes a bicycle to have been present in the camera's field of view. Bicycle's have the advantage in that they are a commonly seen outdoor object, often seen side-by-side with pedestrians.

The design assumes that all bicycle wheels are 622mm in diameter. Occasionally a bicycle wheel may be detected which is far larger or smaller than any surrounding detected bicycle wheel. Ideally, the function will need to detect and ignore outliers. If the system encounters a bicycle under this (for example a child's bicycle) it will treat it like any other bicycle and it would be seen in the inaccurate results. This could be remedied by adding extra complexity to the system. A method of doing so would be to check the expected height of the bicycle wheel at that point. If the detected and expected are not similar the detected height is discarded. This of course will only work if the newly found wheel is within a triangle of points where correct wheels were found.

To be able to calculate a moving object's real world height would help solve the occlusion problem, if their real world heights were calculated both before and after occlusion, these heights could be matched so the tracking system could assign them correct labels.

In a multiple camera surveillance system, tracking people between fields of view is a problem. It is difficult to re-identify people as they walk from one camera's field of view to another.

This dissertation aims to solve these problems. By getting heights before occlusion and then after, the system could decide who is who by matching the calculated heights. But what if the heights could not be calculated before and/or after?

The same problem presents itself for moving between different camera's fields of view.

Another identified problem is in the way people walk. Human walking involves rhythmic up-and-down displacement of the upper body (from pelvis to head), hence the apparent bobbing of the head [Inman et al., 1981] [Rose et al., 2006]. Furthermore, these vertical movements must occur in a smooth sinusoidal manner for the conservation of energy [Inman et al., 1981]. This movement can have an effect of the apparent height of the person; their pixel height is not the same depending on what stage they are in in the walking cycle.

3.4 Conclusion

The goal of this project is to correctly estimate the height of people on video to be used to aid in people tracking and identification.

An overview of design of the proposed system has been given.

The system first finds all moving objects in the video. Each moving object is identified as either a bicycle, person or unidentified object (neither bicycle nor person). If the moving object was found to be a bicycle its wheels are found and measured in pixels. The pixel height of the wheels are stored as well as each wheel's position. If the moving object was found to be a person, that person's real world height is calculated by first checking if the person is standing on a point where a wheel has been found and then calculating their real world height.

4 Implementation

In this section a more detailed implementation of the system is given.

4.1 Introduction

The goal of this project is to correctly estimate the height of people on video to be used to aid in people tracking and identification.

An overview of design of the proposed system has been given.

The system first finds all moving objects in the video. Each moving object is identified as either a bicycle, person or unidentified object (neither bicycle nor person). If the moving object was found to be a bicycle its wheels are found and measured in pixels. The pixel height of the wheels are stored as well as each wheel's position. If the moving object was found to be a person, that person's real world height is calculated by first checking if the person is standing on a point where a wheel has been found, if it is then the real world height of the person can be calculated.

C++ and OpenCV (a library of computer vision functions) were used for this systems implementation.

4.2 Finding Moving Object

Background subtraction is a common and widely used technique for generating a foreground mask when using static cameras. A foreground mask is a binary image containing the pixels belonging to moving objects in the scene. Background subtraction works by performing a subtraction between the current frame and a background model to make a difference image. The background model contains the static part of the scene or everything that can be considered as the background given the characteristics of the observed scene.

A difference image's non-moving pixels (pixels that are the same as that pixel in the background model) will be black, while moving pixels (pixels that are not the same as the background model) will have a non-black colour. The non-black pixels are considered the foreground. Assuming that both the current frame and the background model are greyscale images, the resulting difference image will also be a greyscale image. In this case, the pixel values of the difference image $d(i,j)$ are simply the absolute value of the difference between the pixel values of the video frame $f(i,j)$ and those of the background image $b(i,j)$. The equation for this is given below:

$$d(i,j) = |(f(i,j) - (b(i,j)|$$

To make a binary (black and white) image, a threshold ε can be used. Assuming again that the current frame and the background model are still greyscale images, if the absolute difference between their pixel values is less than the chosen threshold the difference image pixel is considered to be background and is given a value of 0 (black). Otherwise, the difference image pixel is considered to be foreground and is typically given a value of 255 (white).

$$d(i,j) = \begin{cases} 0 & \text{if} f(i,j) - b(i,j) < \varepsilon \\ 255 & \text{otherwise} \end{cases}$$

The background subtraction technique is widely used and is a fundamental task in numerous applications including the detection, tracking and classification of objects, visual surveillance (e.g. crowd monitoring, people counting and action recognition), the detection of security applications (e.g. abandoned luggage, theft and loitering), video annotation and video forensics.

Background modelling consists of two main steps:

- Background Initialization
- Background Update To maintain a background image with a reasonable degree of accuracy it must be updated as the background of the scene changes. It is important that a background image be updated to adapt

to changes such as an object left in the scene so that they will not still be considered part of the foreground.

To implement this, the Gaussian Mixture Model [Stauffer and Grimson, 1999] was used.

Some objects (for example leaves on a tree in the wind) will be moving constantly but also very slightly. We want to ignore this movement. With Gaussian Mixture Model background modelling, each point of the background is modelled using a mixture of Gaussian distributions (around 3 or 4). Each distribution has a weighting depending on how often it has occurred in the past frame. In new frames, each point is compared to the Gaussian distributions to find if there is one close enough. If a distribution is close enough, it is updated If no distribution is close enough, a new Gaussian distribution is initialized. There is a limit on the number of distributions that can be defined for a pixel and, once this limit is reached, the smallest distribution must be discarded in favour of new ones.

For each point the largest distribution is considered the background.

The openCV implementation of the Gaussian Mixture Model was used, which also attempts to detect shadows.

In a binary difference image, black pixels typically correspond to background while white pixels correspond to foreground. Ideally, the classifications of background and foreground pixels would be completely accurate with the difference images consisting solely of true positives (foreground pixels correctly identified as foreground pixels) and true negatives (background pixels correctly identified as background pixels). In practice, however, a number of false positives (background pixels incorrectly identified as foreground pixels) and false negatives (foreground pixels incorrectly identified as background pixels) will also be present.

The number of misclassified pixels that are present in a difference image

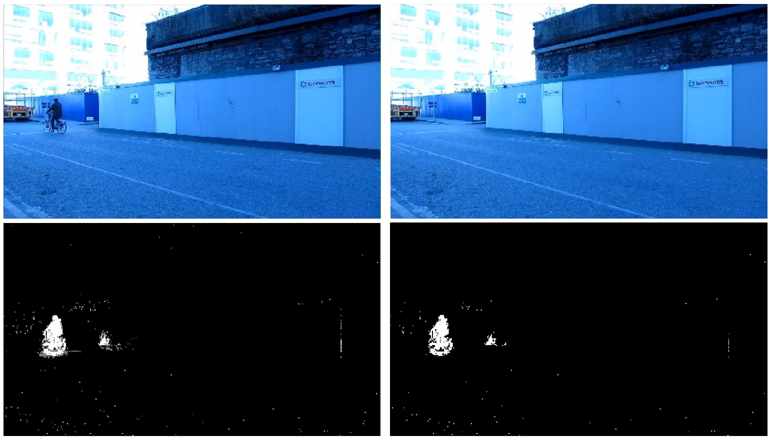

Figure 1: From left: The current frame, background image, foreground mask (showing shadows in grey) and threshold image (removing detected shadows).

should be minimised but, unfortunately, it is not typically possible to completely avoid them. One of the major reasons for this is that background subtraction is dependent upon the background and the foreground of the video frames being distinct or highly contrasting which can generally not be guaranteed. Additionally, it is important that the value of the threshold be set appropriately. If the threshold value is too low then the number of false positives is likely to be significant while, if it is too high, there is likely to be a large number of false negatives.

4.3 Finding People

To find people the HOG (Histogram of Orientated Gradients) People Detector was used.

The persons height in the image is taken to be the bounding box height of the binary silhouette. Note this assumes no errors in the silhouette (such as shadows), no clothing style that modifies the persons apparent height significantly (such as a very long hat), and obviously that the person is walking upright.

4.3.1 Image Gradient

An image gradient is a directional change in the intensity or colour in an image. Image gradients may be used to extract information from images. This gradient of an image measures how it is changing. It provides two pieces of information. The magnitude of the gradient tells us how quickly the image is changing, while the direction of the gradient tells us the direction in which the image is changing most rapidly. Because the gradient has a direction and a magnitude, this information is stored in a vector. Each vector's length provides the magnitude of the gradient, while its direction gives the gradient direction. The image can be represented with a different vector at each point in the image. At each image point, the gradient vector points in the direction of largest possible intensity increase, and the length of the gradient vector corresponds to the rate of change in that direction.

4.3.2 Edges

A method to detect edges in an image can be performed by locating pixel locations where the gradient is higher than its neighbours (or to generalize, higher than a threshold).

Image gradients can be used to extract information from images. Gradient images are created from the original image for this purpose. Each pixel of a gradient image measures the change in intensity of that same point in the original image, in a given direction. To get the full range of direction, gradient images in the x and y directions are computed.

One of the most common uses is in edge detection. After gradient images have been computed, pixels with large gradient values become possible edge pixels. The pixels with the largest gradient values in the direction of the gradient become edge pixels, and edges may be traced in the direction perpendicular to the gradient direction. One example of an edge detection algorithm that uses gradients is the Canny edge detector.

4.3.3 Histogram of Oriented Gradients (HOG)

HOG is a dense feature extraction method for images. Dense means that it extracts features for all locations in the image (or a region of interest in the image) as opposed to only the local neighbourhood of key points. Intuitively it tries to capture the shape of structures in the region by capturing information about gradients. It does so by dividing the image into small cells and blocks of cells. Each cell has a fixed number of gradient orientation bins. Each pixel in the cell votes for a gradient orientation bin with a vote proportional to the gradient magnitude at that pixel.

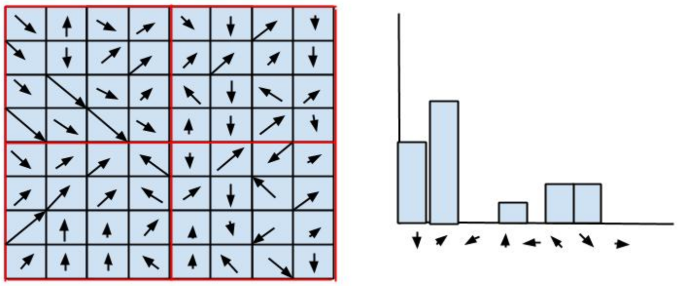

Figure 2: Left shows 4 cells with each pixel vote. Right shows the top right cell's histogram.

To reduce aliasing (which can occur when sampling where the result is distorted), the pixels votes are bi-linearly interpolated. Interpolation happens in both the orientation and position. This means that a pixel will not only vote for its orientation bin, but also for the to neighbouring orientation bins. For example, if the gradient orientation at a pixel is 45 degrees, it will vote with a weight of 0.5 for the 35 to 45 degree bin and a weight of 0.5 for the 45 to 55 degree bin.

Similarly, it will vote for these two orientation bins not only in its cell, but also in the 4 neighbouring cells of its cell. The weights here are decided by the distance of the pixel from the cell centres.

Histograms are also normalized based on their energy (regularized L2 norm) across blocks. Since the blocks have a step size of 1 cell, a cell will be part of 4 blocks. This defines four differently normalized versions of the cell's histogram. These 4 histograms are connected in series to get the descriptor for the cell. Typically, the elements of histograms are also capped at some value.

It then looks in the bin and tries to determine a consensus of slopes to determine where the edges are. HOG can be used to detect any shape including a Person.

4.4 Finding Bicycles

The persons height in the image is taken to be the bounding box height of the binary silhouette. Note this assumes no errors in the silhouette (such as shadows), no clothing style that modifies the persons apparent height significantly (such as a very long hat), and obviously that the person is walking upright.

To find a bicycle, ellipses are found on each side of the subsection of the image which contains the moving object, this continues until an ellipse on each side is found where heights of these ellipses are roughly equal.

4.4.1 Finding Ellipses

An ellipse is found by selecting 5 points all from either the left or right of the section that encompasses the moving object. These 5 points are then used to create an ellipse. If this ellipse is seen to be a good enough fit for the area, it is deemed a possible wheel and the other side is then checked. If the ellipse is not good enough it tries again. There is a maximum number of times to try before the area is deemed to not contain an ellipse (therefore the section does not contain a bicycle wheel and the moving object is not a bicycle).

4.4.1.1 Calculating the Ellipse Bicycle wheels are obviously in the shape of a circle, however the wheels will not always be moving perpendicular to the camera. If the wheel is moving at an angle, it looks like an ellipse.

A conic section is a curve obtained as the intersection of a cone with a plane. Traditionally, the three types of conic section are the hyperbola, the parabola, and the ellipse/circle.

Equation of a conic section [Had2know.com, 2015]:

$$ax^2 + bxy + cy^2 + dx + ey + f = 0 \tag{1}$$

Whether the curve is a hyperbola, parabola, ellipse, or circle depends on the value of the discriminant, $b^2 - 4ac$. The three cases are:

$$b^2 - 4ac > 0, \text{ hyperbola}$$

$$b^2 - 4ac = 0, \text{ parabola}$$

$$b^2 - 4ac < 0, \text{ ellipse or circle (circle only if } b = 0 \text{ and } a = c\text{)}$$

Suppose the coordinates of the five points used to calculate the ellipse are: $(x_0, y_0), (x_1, y_1), (x_2, y_2), (x_3, y_3)$, and (x_4, y_4). The equation of the conic section is found by calculating the determinant of a 6-by-6 matrix and setting it equal to 0. The matrix equation is [Mathworld.wolfram.com, 2015]:

$$\begin{bmatrix} x^2 & xy & y^2 & x & y & 1 \\ x_0^2 & x_0y_0 & y_0^2 & x_0 & y_0 & 1 \\ x_1^2 & x_1y_1 & y_1^2 & x_1 & y_1 & 1 \\ x_2^2 & x_2y_2 & y_2^2 & x_2 & y_2 & 1 \\ x_3^2 & x_3y_3 & y_3^2 & x_3 & y_3 & 1 \\ x_4^2 & x_4y_4 & y_4^2 & x_4 & y_4 & 1 \end{bmatrix} = 0 \tag{2}$$

The coefficients a, b, c, d, e, and f are found from the minors of the the larger matrix [Had2know.com, 2015].

$$a = \begin{bmatrix} x_0y_0 & y_0^2 & x_0 & y_0 & 1 \\ x_1y_1 & y_1^2 & x_1 & y_1 & 1 \\ x_2y_2 & y_2^2 & x_2 & y_2 & 1 \\ x_3y_3 & y_3^2 & x_3 & y_3 & 1 \\ x_4y_4 & y_4^2 & x_4 & y_4 & 1 \end{bmatrix}, b = - \begin{bmatrix} x_0^2 & y_0^2 & x_0 & y_0 & 1 \\ x_1^2 & y_1^2 & x_1 & y_1 & 1 \\ x_2^2 & y_2^2 & x_2 & y_2 & 1 \\ x_3^2 & y_3^2 & x_3 & y_3 & 1 \\ x_4^2 & y_4^2 & x_4 & y_4 & 1 \end{bmatrix} ...etc. \qquad (3)$$

Computing the matrix determinant gives the equation of the ellipse.

The semi-minor (height of the ellipse), semi-major (width of the ellipse) and centre of the ellipse can be found using these coefficients.

Semi-minor Axis [Mathworld.wolfram.com, 2015]:

$$a' = \sqrt{\frac{2(a(\frac{e}{2})^2 + c(\frac{d}{2})^2 + f(\frac{b}{2})^2 - 2\frac{b}{2}\frac{d}{2}\frac{e}{2} - acf)}{((\frac{b}{2})^2 - ac)(\sqrt{(a-c)^2 + 4(\frac{b}{2})^2} - (a+c))}} \qquad (4)$$

Semi-major Axis [Mathworld.wolfram.com, 2015]:

$$b' = \sqrt{\frac{2(a(\frac{e}{2})^2 + c(\frac{d}{2})^2 + f(\frac{b}{2})^2 - 2\frac{b}{2}\frac{d}{2}\frac{e}{2} - acf)}{((\frac{b}{2})^2 - ac)(-\sqrt{(a-c)^2 + 4(\frac{b}{2})^2} - (a+c))}} \qquad (5)$$

The centre (x_{centre}, y_{centre}) of the ellipse is given by [Mathworld.wolfram.com, 2015]:

$$x_{center} = \frac{c\frac{d}{2} - \frac{b}{2}\frac{e}{2}}{(\frac{b}{2})^2 - ac}, y_{center} = \frac{a\frac{e}{2} - \frac{b}{2}\frac{d}{2}}{(\frac{b}{2})^2 - ac} \qquad (6)$$

This ellipse is checked against the shape in the image. If most of the points of the calculated ellipse is accounted for then the program defines that object as an ellipse.

4.5 Conclusion

This section described how idea of using bicycle wheels to find the height of other standing people in the scene was implemented.

The system is made up of a moving object detector (background subtraction), people detector (HOG people detector), and a bicycle wheel detector.

5 Results

Evaluations and tests were performed on the system that was developed. These will be discussed in this chapter with results

In order to test the developed system for this project, a small data set was made.

5.1 Testing Difficulties

Subtle and often lighting changes create difficulties in detecting moving objects for the developed system. To work around this the videos were edited and the learning rate (how quickly it decides a non-moving pixel is part of the background) of the background model (discussed in Section 4.2) was increased. Increasing the background model learning rate also made it more difficult to find the correct size of moving objects.

It was difficult to strike a balance between the two. A limited number of videos from the test dataset were usable.

5.2 Tables

A small test data set was made with two pedestrians and one cyclist. This was to test various real-world scenarios.

5.2.1 Test Video 1

The first video was tested where height of the pedestrians were only calculated if they moved to a point where a wheel had been found.

	Actual Height	Median Height	Mean Height
Person A	1.72m	1.728m	1.75m
Person B	1.85m	1.84m	1.829m

Height calculations between two points where a wheel was detected. This also allowed for when the person was standing on a point a wheel had been detected. If the person was not standing on a point where a wheel had been detected then they system would check if the person was standing on a line between two points where wheels had been found.

	Actual Height	Median Height	Mean Height
Person A	1.72m	1.73m	1.74m
Person B	1.85m	1.71m	1.67m

Height calculations between two points where a wheel was detected. If the person was not standing on a point where a wheel was detected or between two, then it was checked if it was within a triangle of points.

	Actual Height	Median Height	Mean Height
Person A	1.72m	1.7m	1.61m
Person B	1.85m	1.71m	1.71m

5.2.2 Test Video 2

The second video had difficulty distinguishing between the background and foreground due to constant lighting changes in the scene.

where height of the pedestrians were only calculated if they moved to a point where a wheel had been found.

	Actual Height	Median Height	Mean Height
Person A	1.72m	1.76	1.76
Person B	1.78m	2.16	2.17

5.3 Evaluation

The system works with low-resolution images of people, and is robust to changes in clothing.

The first video had no problems with lighting changes and the system performed well. It can be seen in the results that a person's height is more correctly calculated when they are standing on a point where a wheel had been detected rather than between points.

The second video had difficulty distinguishing moving objects from the background and the developed system did not perform well. In the first case (where a person's height was only calculated if they were standing on a point where a wheel had been found), there were a limited number of calculated heights.

5.3.1 Possible Issues

Reasons for error in the developed system are given below as well as ethical issues that might be raised regarding an autonomous video surveillance system.

5.3.1.1 Issues with Implementation The persons height in the image is taken to be the bounding box height of the binary silhouette. Note this assumes no segmentation errors in the silhouette (such as shadows), no clothing style that modifies the persons apparent height significantly (such as a very long hat), and obviously that the person is walking upright.

Shadows cast by moving objects, meanwhile, present significantly more challenging issues. To a background model, such shadows will appear to be moving objects and may be considered part of the foreground. It can be very difficult to distinguish between a moving object and its shadow. For example, a shadow detected as foreground may alter the shape of a true foreground object complicating its detection or, in this application which is concerned with people's heights, a persons shadow may add to the height of the person. Another issue seen was that wheels were sometimes detected in shadows when there was no wheel present.

Confusion also arises when moving objects stop moving. It is unclear whether an object should immediately become part of the scene background once it has come to a stop, whether it should be integrated into the background after a period of time or whether it should ever become part of the background. If a moving person is the object of interest in a scene, for example, it is likely that, if they come to a stop, they will remain of interest, but the system may begin to incorporate them into the background.

Different weather conditions in an outdoor scene can also create considerable difficulties. Wind can contribute to the presence of a dynamic background. It could also cause a camera to shake making for an unsteady video stream. Rain tends to darken the scene, create noise and can alter large portions of the scene, for example colour changes in some materials when wet. Falling snow would be mistaken for foreground and would brighten the scene. Fog will reduce the contrast of the background and foreground in a scene making them more difficult to distinguish. If the sun suddenly becomes occluded by clouds (of the opposite), illumination will change and the scene will suddenly become darker (or brighter).

The inner rim of a bicycle wheel is 622mm [Brown and Allen, 2014]. The detector used in this test system would often find the outer rim, or somewhere between the two.

5.3.1.2 Ethical Issues An important factor especially for large scale applications is the willingness of people to accept the wide deployment of biometric systems. Video surveillance infrastructures might violate a citizens right to anonymity and invade his/her privacy. For example in the US there is resistance to fingerprints because of criminal connotations. There is also the perception of ”Big Brother is Watching you”. These concerns are even more pronounced due to the unobtrusiveness of the height capturing process, which could allow continuous monitoring of people in public places. Apart from this, there is a growing concern that biometrics might be used for purposes beyond the original scope or that unauthorized persons may gain access to biometric information and use it for unlawful purposes. There is also the fear that an innocent person might be banned or held and questioned due to the results of the system.

6 Conclusion

The goal of this project was to correctly estimate the height of people on video to be used to aid in people tracking and identification.

The approach of this dissertation was to use an object of a known size to estimate height of standing people in a scene to better track and re-identify them as they move between fields of view. The system would be able to calculate the size of any detected object in the scene by comparing it to the size of the known object which would normally be seen in the scene. This object would be a bicycle wheel.

A system was developed to test this using C++ and OpenCV. The developed system depends on lighting not changing in the video.

Though much more testing is needed, it can be seen a person's height is more correctly calculated when they are standing on a point where a wheel had been detected rather than between points.

6.1 Future Work

While the focus on this dissertation was to help identify people in a video surveillance network with an emphasis on security, the person height information can contribute to other applications such as marketing.

Perhaps the best approach for achieving practical identification results is to combine this calculated height with other biometrics, such as what was discussed in section 2.

Height could provide context when it is considered along with other biometrics to correctly estimate the age and gender of a person in a scene.

Future work should also aim to develop and incorporate more robust methods of bicycle wheel detection, people detection and moving object detection or investigate improved methods. These improved methods should produce more stable results.

Further experiments on bigger database should be carried out to better test the proposed method.

References

[Brown and Allen, 2014] Brown, S. and Allen, J. (2014). Measuring bicycle rims and hub flanges.

[Chen et al., 2001] Chen, J., Zhang, C., and Rong, G. (2001). Palmprint recognition using crease. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 3, pages 234–237. IEEE.

[Chorás, 2007] Chorás, M. (2007). Emerging methods of biometrics human identification. In *Innovative Computing, Information and Control, 2007. ICICIC'07. Second International Conference on*, pages 365–365. IEEE.

[Criminisi et al., 1999] Criminisi, A., Reid, I., and Zisserman, A. (1999). Single view metrology. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 1, pages 434–441 vol.1.

[Dalal and Triggs, 2005] Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:886–893.

[Gallagher et al., 2009] Gallagher, A., Blose, A. C., and Chen, T. (2009). Jointly estimating demographics and height with a calibrated camera. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1187–1194.

[Gomez et al., 2002] Gomez, E., Travieso, C. M., Briceno, J., and Ferrer, M. (2002). Biometric identification system by lip shape. In *Security Technology, 2002. Proceedings. 36th Annual 2002 International Carnahan Conference on*, pages 39–42. IEEE.

[Had2know.com, 2015] Had2know.com (2015). How to find the equation of a conic section through five points.

[Hoiem et al., 2008] Hoiem, D., Efros, A. A., and Hebert, M. (2008). Putting objects in perspective. *International Journal of Computer Vision*, 80(1):3–15.

[Hoogstrate et al., 2001] Hoogstrate, A., Van den Heuvel, H., and Huyben, E. (2001). Ear identification based on surveillance camera images. *Science & Justice*, 41(3):167–172.

[Hu et al., 2004] Hu, W., Tan, T., Wang, L., and Maybank, S. (2004). A survey on visual surveillance of object motion and behaviors. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 34(3):334–352.

[Iannarelli, 1989] Iannarelli, A. V. (1989). *Ear identification.* Paramont Publishing Company.

[Inman et al., 1981] Inman, V. T., Ralston, H. J., and Todd, F. (1981). *Human walking.* Williams & Wilkins.

[Kasprzak, 2003] Kasprzak, J. (2003). Polish methods of earprint identification. In *Forensic Science International*, volume 136, pages 349–349. Elsevier Sci Ireland Ltd.

[Khan and Shah, 2003] Khan, S. and Shah, M. (2003). Consistent labeling of tracked objects in multiple cameras with overlapping fields of view. *IEEE Trans. Pattern Anal. Machine Intell.*, 25(10):1355–1360.

[Kispál and Jeges, 2008] Kispál, I. and Jeges, E. (2008). Human height estimation using a calibrated camera. *Proc. of the computer vision and pattern recognition (CVPR 2008).*

[Lalonde et al., 2007] Lalonde, J.-F., Hoiem, D., Efros, A. A., Rother, C., Winn, J., and Criminisi, A. (2007). Photo clip art. In *ACM Transactions on Graphics (TOG)*, volume 26, page 3. ACM.

[Mathworld.wolfram.com, 2015] Mathworld.wolfram.com (2015). Ellipse – from wolfram mathworld.

[Nixon et al., 1999] Nixon, M., Carter, J., Nash, J., Huang, P., Cunado, D., and Stevenage, S. (1999). Automatic gait recognition. In *Motion Analysis and Tracking (Ref. No. 1999/103), IEE Colloquium on*, pages 3/1–3/6.

[Nixon and Carter, 2004] Nixon, M. S. and Carter, J. N. (2004). Advances in automatic gait recognition. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 139–144. IEEE.

[Pasescu and Tanislav, 1997] Pasescu, G. and Tanislav, E. (1997). Person identification on the basis of earprints in the activity of bucharest police department. *Information Bulletin for SP/TM Examiners*, 3(3).

[Phillips et al., 1998] Phillips, P., McCabe, R., and Chellappa, R. (1998). Biometric image processing and recognition. In *Signal Processing Conference (EUSIPCO 1998), 9th European*, pages 1–8.

[Rose et al., 2006] Rose, J., Gamble, J. G., and Adams, J. M. (2006). *Human walking*. Lippincott Williams & Wilkins Philadelphia.

[Stauffer and Grimson, 1999] Stauffer, C. and Grimson, W. (1999). Adaptive background mixture models for real-time tracking. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:246–252.

[Topcu et al., 2014] Topcu, O., Alatan, A. A., and Ercan, A. O. (2014). Occlusion-aware 3d multiple object tracker with two cameras for visual surveillance. *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 56 – 61.

[Vos, 2015] Vos, T. (2015). *The Video Surveillance Industry Market size*. www.cameramanager.com, 1 edition.

[Yesil, 2006] Yesil, B. (2006). Watching ourselves. *Cultural Studies*, 20(4-5):400–416.