Applying Computer Analysis to Detect and Predict Violent Crime during Night Time Economy Hours.

A thesis submitted in partial fulfilment

of the requirement for the degree of Doctor of Philosophy

Kaelon Lloyd

July 2018

Cardiff University
School of Computer Science & Informatics

Declaration

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed .... (candidate)

Date ....

Statement 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD.

Signed .... (candidate)

Date ....

Statement 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references.

Signed .... (candidate)

Date ....

Statement 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .... (candidate)

Date ....

Abstract

The Night-Time Economy is characterised by increased levels of drunkenness, disorderly behaviour and assault-related injury. The annual cost associated with violent incidents is approximately £14 billion, with the cost of violence with injury costing approximately 6.6 times more than violence without injury. The severity of an injury can be reduced by intervening in the incident as soon as possible. Both understanding where violence occurs and detecting incidents can result in quicker intervention through effective police resource deployment.

Current systems of detection use human operators whose detection ability is poor in typical surveillance environments. This is used as motivation for the development of computer vision-based detection systems. Alternatively, a predictive model can estimate where violence is likely to occur to help law enforcement with the tactical deployment of resources. Many studies have simulated pedestrian movement through an environment to inform environmental design to minimise negative outcomes. For the main contributions of this thesis, computer vision analysis and agent-based modelling are utilised to develop methods for the detection and prediction of violent behaviour respectively.

Two methods of violent behaviour detection from video data are presented. Treating violence detection as a classification task, each method reports state-of-the-art classification performance and real-time performance. The first method targets crowd violence by encoding crowd motion using temporal summaries of Grey Level

Co-occurrence Matrix (GLCM) derived features. The second method, aimed at detecting one-on-one violence, operates by locating and subsequently describing regions of interest based on motion characteristics associated with violent behaviour. Justified using existing literature, the characteristics are high acceleration, non-linear movement and convergent motion. Each violence detection method is used to evaluate the intrinsic properties of violent behaviour.

We demonstrate issues associated with violent behaviour datasets by showing that state-of-the-art classification is achievable by exploiting data bias, highlighting potential failure points for feature representation learning schemes.

Using agent-based modelling techniques and regression analysis, we discovered that including the effects of alcohol when simulating behaviour within city centre environments produces a more accurate model for predicting violent behaviour.

Acknowledgements

Firstly, I would like to express my sincere gratitude to Prof. Dave Marshall for providing the research opportunities that culminated in the development of my thesis. Along with Prof. Marshall, I would also like to extend my gratitude to my secondary and tertiary supervisors Prof. Paul Rosin and Prof. Simon Moore for their support of my PhD study. Their insightful comments and questions helped me to widen my research and develop lifelong skills.

I would like to thank my peers for providing stimulating discussions and for all the fun we have shared over the last few years. I am especially grateful to the constituents of C2.13, both the old and new. You created an enjoyable environment which was much needed during my time writing and finalising my thesis. Specifically, I am grateful to Matthew Nunes for allowing me to experience his singular vocal expression and for engaging me in his risqué conversations.

To Nyala Noe, you impacted my progress like no other. You helped me regain the motivation to work after a period of hardship. Without your positive influence, the prospect of submission seems less likely.

I would like to thank my family and friends whose support over the years has been invaluable. Finally, I would like to thank various Tea production companies. Your product was an integral component of the beverages that have kept me hydrated over the years.

Contents

List of Tables

List of Acronyms

ABM Agent Based Modelling

AIC Akaike Information Criterion

AMV Acceleration Measure Vector

AT Assistive Technologies

AUC Area Under Curve

AUROC Area Under Receiver Operating Characteristic Curve

BoW Bag-of-Words

BAC Blood Alcohol Concentration

CA Cellular Automata

CAA Computer Aided Analysis

CAD Computer Aided Diagnosis

CAO Computer Aided Observation

CCTV Closed Circuit Television

CNN Convolutional Neural Network

CRP Crime Reduction Programme

CVPP Cardiff Violence Prevention Program

DoG Difference of Gaussians

EM Expectation Maximisation

FPS Frames Per-Second

FTLE Finite Time Lyapunov Exponent

GCF Global Contrast Factor

GCFM Generalized Centrifugal Force Model

GLCM Grey Level Co-occurrence Matrix

GMM Gaussian Mixture Model

GPS Global Positioning System

HOFO Histogram of Optical Flow Orientation

IDT Improved Dense Trajectories

IFU Inter-Frame Uniformity

ILF Inverse Laminar Flow

IFV Improved Fisher Vector

IQA Image Quality Assessment

JPEG Joint Photographic Experts Group

KDE Kernel Density Estimation

KLT Kanade–Lucas–Tomasi

LDA Latent Dirichlet Allocation

LSTM Long Short Term Memory

MLV Maximum Local Variation

MOHMM Multi-Observation Hidden Markov Model

MoSIFT Motion Scale-Invariant Feature Transform

NPPV Non Police Personnel Vetting

NTE Night-Time Economy

OLS Ordinary Least Squares

OViF Oriented Violent Flows

PAI Predictive Accuracy Index

PCA Principal Component Analysis

PSI Personal Space Invasion

PSNR Peak Signal-to-Noise Ratio

RIMOC Rotation Invariant Motion Coherence

RMSE Root Mean Square Error

ROC Receiver Operating Characteristic

SFM Social Force Model

SGD Stochastic Gradient Descent

SIFT Scale-Invariant Feature Transform

SVM Support Vector Machine

UK United Kingdom

VAP Violence Against Person

ViF Violent Flows

Chapter 1

Introduction and Motivation

1.1 Motivation

The work presented within this thesis was motivated by observations of the Night-Time Economy (NTE) within the United Kingdom (UK). In the context of this work, the NTE refers the economic and social activities that occur between the hours of 6pm and 6am. Increased levels of drunkenness, disorderly behaviour and assault-related injury characterise the NTE in Britain [44, 83]. The Crime Survey for England and Wales documented an estimated number of 1.2 million violent incidents for the year ending in June 2017, with approximately 46% of those incidents including injury [36]; these figures relate to incidents within and outside the NTE. The annual cost associated with violent incidents is approximately £14 billion, with the cost of violence with injury costing approximately 6.6 times more than violence without injury [34]. Identifying mechanisms for reducing *violence with injury* to *violence without injury* could result in great economic benefit as well as providing benefits to the victims of violence. Ultimately, the main goal of violence research would be to determine mechanisms that not only reduce the severity of an incident, but also stops violent acts from occurring.

In 1997, a Cardiff based multi-agency violence prevention group was formed. The group members coalesced to form the Cardiff Violence Prevention Program (CVPP) which launched in 2003. The primary focus of the CVPP was the utilisation of a data sharing network that allows for the sharing of anonymised data collected from police, health and local government services. The gathered information was used to devise crime prevention tactics and strategies with the intent of reducing violent crimes. The founders of the CVPP validate their crime prevention methods by comparing rates of crime with 14 cities identified as being statistically similar to Cardiff. The study concluded in 2007 and showed that violent crime prevention measures resulted in an estimated 42% fewer woundings [37] with a follow-up paper on the economic effects of this program suggesting saving an estimated £6.9 million in 2007, with a £1.25 million reduction in health service costs [38]. This research demonstrates the economic impact of violence reduction strategies. However, a review of the literature suggests sub-optimal use of technology when attempting to reduce violence and its effects.

Video Surveillance and Violence

The research developed by the CVPP applied video surveillance systems as part of their crime prevention strategies. Cameras were used to monitor violent hotspots around the city centre in an active capacity, and from this, the authors hypothesise that monitoring for violence allows for faster police deployment and subsequent crime intervention. This hypothesis stems from the observation that injury severity decreased after applying a policing strategy that decreased intervention time [38]. These results corroborate Sivarajasingam *et al.* [127] who present a hypothesis that Closed Circuit Television (CCTV) plays an integral role in reducing violent crime intervention time and therefore reduces the severity of the incident.

Violence in the context of this thesis refers to situations in which the behaviour of one or more persons inflicts physical bodily harm to another person. Violent behaviour during NTE hours is typically characterised through the use of close proximity actions like punching, kicking, pushing, jostling or grappling. Incidents that involve indirect interactions like throwing projectiles are unexamined in this work. Aggression is an associated concept whereby a person behaves with intent to cause physical or emotional harm to another person; aggressive behaviour typically precedes violence, behaviour that concerns the infliction of harm.

Multiple studies found that the installation of CCTV is associated with an increase in violent crime [38, 43, 127]. The consensus is that the installation of CCTV allows for the identification of violent instances that would have gone unnoticed when no cameras were present. By analysing emergency department statistics, existing research [50] has shown the percentage of violent incidents not documented by the police for male victims is 57.4% and 60.3% for street assaults and assaults in licensed premises respectively.

High profile criminal cases and governmental mandates led to the mass adoption of video surveillance systems across the United Kingdom [39]. The estimated number of cameras installed across the United Kingdom falls within the range 1.85 million [105] and 4.2 million [42]. Although the adoption of CCTV systems has shown to lead to an increase in violent crime detection [38, 43, 127], many instances still go unobserved; this is partly due to inadequate human observation ability [30, 51, 100, 109, 127, 133]. Concisely, there is too much information to observe. The simultaneous observation of visual feeds result in reduced observation ability when attempting to locate scenes of interest [133, 139]. Other factors such as the length of time observing [30, 51, 100, 109] and the influence of personal beliefs [47, 106] affect the focus of a human observer.

Computer-based analysis of information is used to augment human capabilities to aid the disabled [78], or improve task completion accuracy and efficiency [31, 129, 130]. The effectiveness at completing many different tasks has improved through the aid of computer-based analysis, for instance, Swett *et al.* [130, 129] describes a system that allows a radiologist to operate at an enhanced level by providing feedback concerning the current working diagnosis based on both patient and historical data. Computer Aided Diagnosis (CAD) has been widely adopted in the medical profession with CAD systems resulting in an increased detection rate of lung, breast and colon cancers [31]. Outside of the medical field, Sarshar *et al.* [118] propose a method used to aid with the inspection of sewer systems by using computer vision to identify defects captured using CCTV for a supervisor to consider when evaluating the condition of a sewer segment. Ratings are used to assess the remaining service life of a sewer system, information that is useful for efficient town/city maintenance and operation. All the methods mentioned have demonstrated an improvement in human efficiency at completing tasks when using an assistive computer vision system. The development of an assistive system has the potential to improve violence crime detection rates using CCTV.

Pedestrian Simulation and Historical Crime Data

The environment and context of a person affect how they behave. Understanding the effects that induce negative behaviour or outcomes can be used to inform the development of preventative strategies. Many studies have simulated pedestrian movement through an environment to inform environmental design to minimise negative outcomes [19, 153, 158]. Models have been utilised to determine how crime is affected by the alteration of the environment [119] and policing strategies [33]. The use of historical data can reveal regions within an environment where harmful behaviour is likely to occur. In the context of crime, identifying these regions and deploying police resources has shown to be beneficial in reducing crime and its effects [10, 122, 123]. Utilising historical crime data can facilitate the development of a predictive model that can then be used to evaluate where crime would occur within a new, unseen environment. The authors of SimDrink present a model for testing the effects of policy changes on violence during NTE hours [119]. Predictive tools will allow law enforcement and city management to deploy strategies that reduce the impact of violent behaviour.

1.2 Research Hypothesis

The general hypothesis for this thesis is as follows:

> Using computer-based analysis, violent crime can be detected and predicted such that the detection time is reduced, and that instances that would typically go undetected, are identified.

Influenced by the data available, two focused hypotheses were derived from the general hypothesis. The hypotheses are as follows:

- Computer vision analysis and machine learning can be used to detect whether video footage depicts violent or non-violent behaviour.
- Pedestrian simulation and data modelling techniques can be used to generate an accurate geo-spatial prediction of violent crime.

The literature suggests that decreasing violence detection time and increasing detection rate will have positive effects on the severity of injury sustained from a violent incident. Before this can be investigated, the general research hypothesis presented above must be proven. Developing the tools required to perform *computer based analysis* to *detect* and *predict* violent crime form the main contributions of this thesis.

1.3 Contributions

Research presented in this thesis provides contribution to two distinct fields, *Computer Vision based Violence Analysis* and *Pedestrian Modelling.* Computer Vision and Pedestrian Modelling concern the detection and prediction of violent behaviour respectively.

Computer Vision: Violent Crowds

This work presents two distinct methods of violent behaviour analysis. Early research suggested that the task of violent behaviour detection could be decomposed into two classes, *One-on-one violence* and *Crowd Violence*. The method outlined in Chapter 4 concerns the detection of violent behaviour within crowded environments. Violence in crowds is characterised by the number of pedestrians involved both directly and indirectly (Figure 1.1). Increased population density affects the perception of violence as individual actions cannot be reliably identified due to occlusions. It is also assumed that dense populations and increased combatant count result in perceived characteristics of violence that differ when compared to one-on-one violence between two people (Figure 1.2).

The research disclosed in Chapter 4 takes inspiration and methodologies from *crowd counting* techniques, in which the number of pedestrians depicted in a single frame is estimated. Existing research has demonstrated that measurements derived from a Grey Level Co-occurrence Matrix (GLCM) [54] are powerful at capturing structural information of a crowd, resulting in high crowd counting accuracy [3, 16, 21, 91, 142]. It is hypothesised that the temporal analysis of GLCM features could be used to capture crowd motion dynamics that are capable of discriminating between different behavioural classes when used in conjunction with a machine learning classifier. Temporal changes in appearance are represented by measuring how GLCM features change over time using four statistics: *mean*, *standard deviation*, *skewness* and *Inter-Frame Uniformity (IFU)*. The method proposed demonstrated state-of-the-art performance when compared to existing real-time violence detection techniques. The contributions of this work are:

1. IFU was designed to measure the stability of appearance over time. Using

Figure 1.1: Still images depicting violent crowds.

three datasets that contain both violent and non-violent samples, it was found that the appearance of violence is less stable/linear over time when compared to similar scenes depicting normality.

2. Analysing longer sections of a video resulted in greater discrimination between violent and non-violent situations.

3. The computational cost of generating GLCMs and their derived measurements is low, allowing for the production of a real-time system, processing data at a rate greater than 30 Frames Per-Second (FPS).

Computer Vision: One-on-One Violence

One-on-one violence in the context of this work is defined as physical and violent engagement between two participants. This type of violence is mainly differentiated from crowd violence by the number of people involved. Typically, with fewer people involved, the perception of violent actions such as punching and kicking are more clear due to reduced visual obstruction caused by the crowd (Figure 1.2).

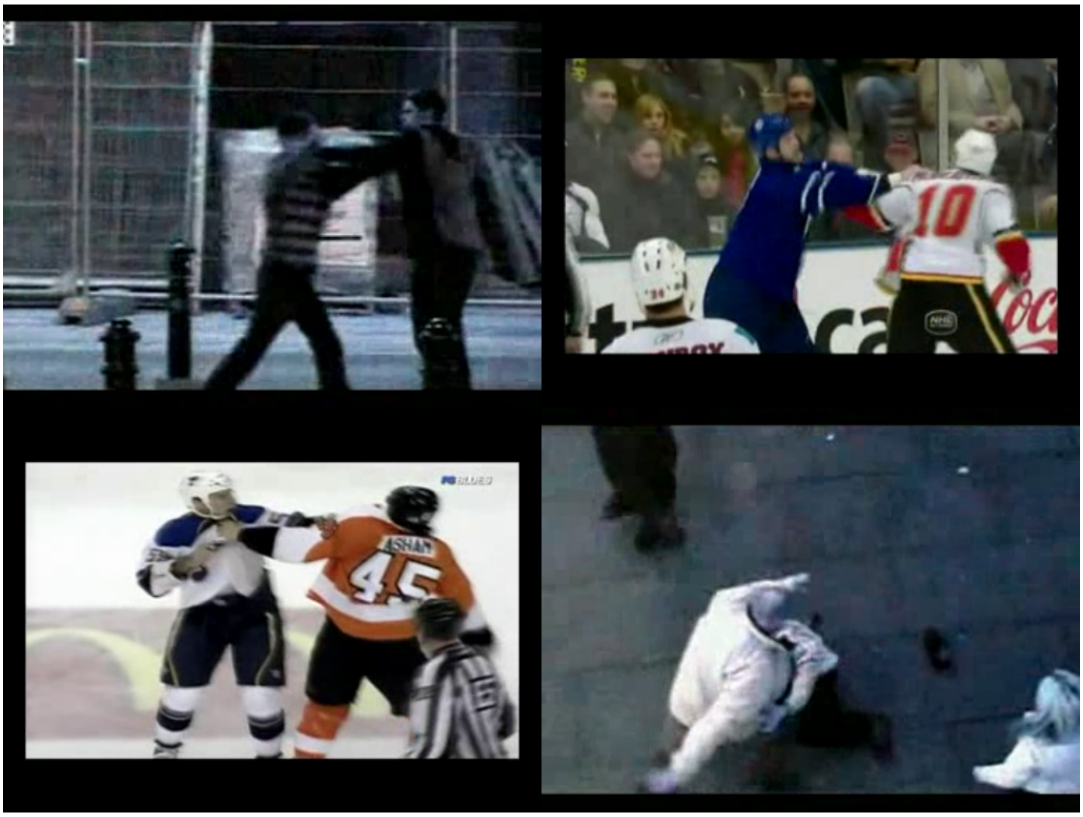


Figure 1.2: Still images depicting one-on-one violence.

Chapter 5 presents an interest-point based detection algorithm for analysing violent behaviour. The presented interest-point based method identifies spatiotemporal subregions in a video volume that contain objects whose kinetic properties are reflective of violent behaviour. This method was initially designed to detect violence that takes the form of one-on-one fighting (a category of violence unsuited by a crowd violence detector), however testing revealed this approach to be suitable towards the analysis of crowds. To briefly summarise, the proposed method computes dense motion trajectories using a process referred to as *particle advection*. The particle advection process extracts accurate motion trajectories from noisy data, such as crowds [98, 102, 162]. Three measurements are extracted from the motion trajectories and used to generate a response map where high values are indicative of regions that have a high violence potential relative to the rest of the scene; interest points are identified using a Difference of Gaussians (DoG) interest point detector [81]. The three measurements, *acceleration*, *convergence*, and *linearity* were designed and justified using existing literature from the fields of Computer Vision [26, 29, 97] and Pedestrian Modelling [59, 156]. Multiple feature vectors are generated using data located around detected interest points identified within a spatiotemporal volume. A spatiotemporal volume is then represented using a set of features that are aggregated using a Bag-of-Words (BoW) model. The following contributions have arisen from this method:

1. Experimentally demonstrated that dense feature sampling provides greater classification ability when attempting to discriminate between violent and non-violent samples.

2. A comparison of feature sampling methods showed that using an interest-point based detector to determine regions for description outperformed a dense grid-based sampling approach for the task of violent behaviour detection.

3. It was demonstrated that the environment or context of violence could alter the nature of the violent interaction. Regarding the underlying dynamics of motion, violence in city centre locations is different than violence that takes place in crowds, and violence found in the sport of ice hockey.

4. Using a one-class learning methodology, it was demonstrated that the proposed method was capable of distinguishing between normal and violent

scenes even when training a classifier without data that represented violent behaviour. Furthermore, the results suggest that one-class learning works better on datasets whose environments and recording equipment are consistent.

Computer Vision: Depth

When recording violent incidents, it is often the case that the camera operator attempts to get as close as possible for the best view. Capturing footage of normal behaviour must be performed similarly. Otherwise, any comparison between the normal and violent behaviour will be biased based on perceived depth. In Chapter 3, the rate at which each violent behaviour dataset used throughout this thesis is affected by the perceived depth bias is quantified. It was demonstrated that publicly available datasets are subject to perceived depth bias. This information is used to inform a discussion regarding the importance of scale invariant or depth corrected analysis.

Agent Based Modelling and Violent Hotspot Analysis

Understanding where and when disorder and violence will occur is of value to efforts aimed at mitigating harm [131]. It is assumed that the physical and social environment influence harm in the NTE through factors associated with population density, street congestion, drinking establishment opening times, and blood alcohol levels [149]. Agent-based models allow for the simulation of real-world pedestrian flow. In many night-time locations, alcohol is used as a social lubricant and is synonymous with an unsteady gait and violence. Empirically justified effects of alcohol, by time and dose, are incorporated in an agent-based model to simulate drunken behaviour [101]. The key behavioural characteristic associated with alcohol induced intoxication are changes in a gait [28, 68, 132, 110]. A person's ability to maintain a normal walking cycle without staggering decreases as they ingest more alcohol, becoming more intoxicated.

The density, velocity and invasions of personal space of simulated pedestrians are measured and used as variables for predicting violent crime hotspots. Violent crime hotspots are generated using a Gaussian Kernel Density Estimation (KDE) process applied to geo-coded crime data obtained from law enforcement agencies; the local spatial distribution of violent crime is assumed to be Gaussian as is typical of hotspot analysis [15, 23, 48, 141]. Regression analysis and statistical correlation testing are used to measure and compare the predictive ability of variables derived from pedestrian simulation using real-world environments and data.

1. Introduction of a *Personal Space Invasion (PSI)* measurement as informed by existing literature on the proximity of human comfort zones [52]. The investigation of PSI was justified by the assumption that invasions of personal space induces violent behaviour. This assumption was derived from literature on stress induced by close proximity interactions [69, 128], and the relationship between stress and violence [4, 22, 73, 95, 103].

2. PSI is a powerful predictor of violent behaviour, providing evidence for the hypothesis that stress-induced violence results from invasions of personal space.

3. Demonstrably shown using regression and statistical testing that an agent-based model informed by characteristics associated with pedestrian intoxication allows for an increased level of violent crime prediction when considered alongside a model that lacks intoxicated characteristics.

4. Presented a foundation for future work for the creation of a more accurate simulation of intoxicated pedestrian behaviour.

1.3.1 Contributions Relative to the Hypotheses

The computer vision hypothesis targets the investigation of an algorithms ability to detect violent behaviour captured on video. The computer vision hypothesis was derived from the general hypothesis and aims to reduce the effects of violence through the detection of violent behaviour. As discussed in Chapter 1 and Chapter 2, the power of reducing the effects of violence through CCTV lies in aiding real-time detection of violence. For this reason, as implied by the general hypothesis, the investigation of the computer vision hypothesis has a constraint of real-time computation attached. Algorithms that have a computational delay are not suited for deployment in an active surveillance capacity. With this consideration in mind, Chapters 4 and 5 present two independent methods of violent behaviour detection that operate in real-time, where real-time is defined as operating at over 30 frames per second. The classification performance of each method for the binary classification task of violent behaviour detection was demonstrated to be non-random (Receiver Operating Characteristic (ROC) > 0.5, Figure 1.3); better than random classification performance is evidence against the rejection of the computer vision hypothesis. Although the reported classification performance is high, which provides evidence for the hypothesis, the criteria for accepting the hypothesis has yet to be developed. Developing the criteria for accepting the violent behaviour detection hypothesis is discussed in Chapter 7, Section 7.5. An investigation into the image quality and perceived depth biases have revealed potential issues with data commonly used to evaluate violent behaviour prediction. Information about data biases, feature sampling, and insight derived from one-class learning, can be used to influence future research that works towards proving the violent behaviour detection hypothesis.

The goal of the prediction hypothesis is to investigate the ability of pedestrian simulation techniques for violent crime prediction. In terms of the work presented in Chapter 6, evidence refuting the hypothesis would manifest if pedestrian

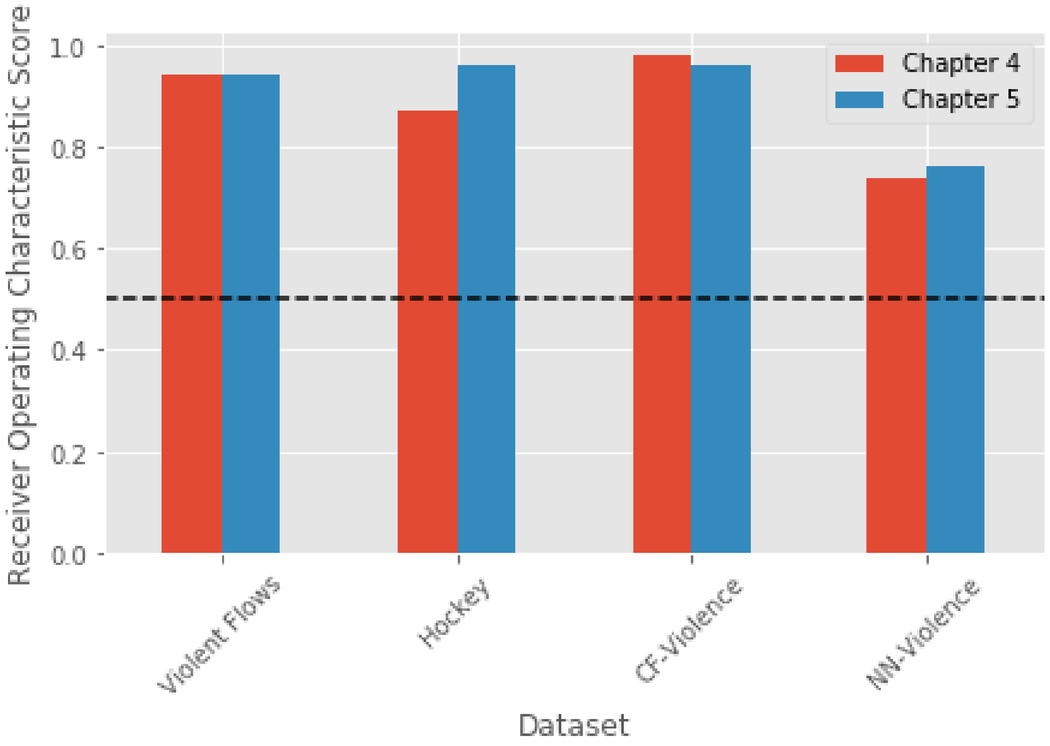


Figure 1.3: Receiver Operating Characteristic score reported by each algorithm presented in this thesis on four different violent behaviour datasets.

simulation produced measures that either did not correlate with violent crime or resulted in a regression model with a low R^2 score. Through correlation analysis and regression modelling, the results in Chapter 6 show that this is not the case. The results in Chapter 6 provides evidence for the hypothesis being true. The introduction of a measure of personal space invasion proved useful for improving regression model performance when compared to a model that considered only pedestrian density and velocity. Additionally, the inclusion of intoxicated processes that simulate drunken behaviour was demonstrated to yield a further increase in violent crime prediction accuracy. Lastly, a theoretical model for improving the intoxicated pedestrian model is proposed in Chapter 6, but not implemented due to data limitations. The threshold for absolute acceptance of the prediction hypothesis is not currently known and needs to be determined. One avenue for determining a threshold of acceptance is to survey expert opinion to discuss the minimum model accuracy required to be usable in real-world conditions. To satisfy the global hypothesis, a controlled study would be required to understand the effects of deploying a prediction system.

1.4 Organisation of Thesis

Chapter 2, Section 2.1 provides a review of the literature associated with the effectiveness of human CCTV observation and computer vision based methods for violent behaviour analysis. Chapter 2, Section 2.4, presents a review of the literature regarding pedestrian simulation, crime prediction, and crime hotspot analysis. The information presented in Chapter 2 reveals the knowledge gap that the work presented in this thesis aims to fill.

The data used for each experiment in this document is introduced and described in Chapter 3. Chapter 3 presents an evaluation of the biases that exist in each video dataset; issues associated with data used to evaluate violent behaviour detection methods are discussed.

Chapters 4 & 5 present computer vision based solutions for the violent behaviour detection task. Chapter 6 presents an investigation into the efficacy of alcohol informed pedestrian simulation model for the task of violent behaviour detection. Finally, Chapter 7 presents conclusions and future work.

1.5 List of Publications

The work introduced in this thesis is based on the following publications.

- Kaelon Lloyd, Paul L. Rosin, David A Marshall, and Simon C. Moore: Detecting violent and abnormal crowd activity using temporal analysis of grey level co-occurrence matrix (GLCM)-based texture measures. *Machine Vision and Applications*, 28(3-4), 361-371, 2017

- Kaelon Lloyd, David A Marshall, Paul L. Rosin, and Simon C. Moore: Violent behaviour detection using local trajectory response. *7th International Conference on Imaging for Crime Detection and Prevention*, 14-6, 2016

Publications Pending

- Kaelon Lloyd, Simon C. Moore, David A Marshall, and Paul L. Rosin: Predicting violent crime using agent-based modelling: Investigating an alcohol informed model, *PlosOne*, 2018 **Under revision**

- Kaelon Lloyd, David A Marshall, Paul L. Rosin, and Simon C. Moore: Violent behaviour detection using local trajectory response:An Analysis Using Real-World CCTV Data. Journal *Pattern Recognition*, 2018, **Undergoing revision**

Chapter 2

Literature Review

2.1 Introduction

Presented in this chapter is a review of all relevant literature to both provide context for, and justify the existence of, research developed for this thesis. The scope of research within this thesis covers two distinct fields of research to solve two related problems, these being the detection and the prediction of violent behaviour. Due to the dual nature of this thesis, the literature review is divided into two sections; the first section concerns surveillance footage analysis and violent behaviour detection. The other section concerns pedestrian behaviour simulation and violent behaviour prediction.

2.2 Video Surveillance and Closed Circuit Television

Closed Circuit Television (CCTV) is a video transmission service that transmits live video feeds through a closed connection to a central point for viewing and archiving. The term CCTV does not, by definition, equate to video surveillance, but it has become heavily associated, resulting in the terms CCTV and *video surveillance systems* being used interchangeably throughout the literature.

It is a widely stated that Britain has the most CCTV cameras per person than any other country with the most common estimate being approximately 4.2 million in total with the attributing fact that a person is seen by 300 cameras per day [105]; a more recent report [42] quotes a figure closer to 1.85 million. A person is expected to fall in view of at most 70 different cameras a day. It is difficult to determine the true CCTV camera count as the United Kingdom Hoeme Office does not keep track of the number of installed CCTV cameras as the task is infeasible due to the number of privately owned surveillance systems [105].

The mass adoption of CCTV in the United Kingdom was fuelled by a few key events in the early 1990s. In particular, an Audit Commission called for "a massive expansion of proactive intelligence led policing and singled out CCTV as having a major role to play in crime prevention [39]". This statement resulted from the observation of criminal activity which was reported in the same document:

> The introduction of closed-circuit TV cameras into Airdrie town centre, with public support, has had dramatic results. In the first twelve months of operation, recorded crimes dropped from 2,475 to 627, of which 447 (71%) were cleared up. Break-ins to commercial premises dropped from 263 to 15 and incidents of vandalism from 207 to 36. The reduced workload in incident response has allowed increased patrol in rural areas. The costs of installation and maintenance are met by the local business community. [39]

The audit report, helped along by the use of CCTV in solving the high-profile murder of James Bulger, greatly affected the speed of adoption of CCTV:

> The extent of the Home Office backing for and reliance on CCTV is indicated by the fact that by 1995, 78% of the Home Office budget for crime prevention was being used to fund schemes to put CCTV in public places. [46]

The United Kingdom government provided funding for CCTV systems until the year 2003 through the Crime Reduction Programme (CRP), at which point the cost associated with installation, use, and maintenance of such systems was transferred to local authorities [41]. Lack of a centralised CCTV management group resulted in local CCTV systems developing differently over time. Early CCTV technology often captured poor quality video footage when compared to today's standards. Unfortunately, discussions with local police organisations revealed that the cost of CCTV upgrades were high, and often cameras have not received an update. As reported by Gerrard *et al.* [41] "Anecdotal evidence suggests that over 80% of the CCTV footage supplied to the police is far from ideal". As informed by the literature, CCTV footage quality is poor in quality and overabundant. The following section will discuss how data overabundance and data quality have an influence on the ability for a human to complete vision-based tasks.

2.2.1 Surveillance and Crime

Surveillance cameras record the actions of objects and people. It is beneficial to identify the actions depicted in order to detect criminal behaviour so that appropriate measures can be performed by security personnel. CCTV can be used either passively or actively. Active CCTV analysis requires a human observer watch a live video feed and identify instances of criminal, or abnormal activity. The active use of CCTV surveillance systems allows the operator to guide the appropriate personnel towards a scene of interest; any developments can be viewed from a strategic location which allows the operator to determine the best course of action to take regarding ground units. Passive usage of CCTV systems manifest itself in one of two ways. The first is archival; when an incident is recorded but not detected, the footage can be retroactively accessed for evaluation or playback during court hearings. The second use of passive CCTV an effect referred to as the

Panopticonisation effect, where the presence of a camera is enough to dissuade people from committing crimes. The term panopticonisation derives from the name given to a prison design by Jeremy Bentham, in which the geometry of the building does not allow prisoners to determine whether they are being observed. Prisoners cannot perform any malicious behaviour with any degree of confidence that they will not be caught. Therefore prisoners cannot confidently assess the risk of their actions and must adopt good behaviour by default. However, the panopticonisation effect does not affect violent incident rates during the NTE [43, 117, 150, 151]. It is hypothesised that violence is a spontaneous activity and that perpetrators do not consider their environment before engaging [43, 117].

> In some cases premeditated, or more planned, offences, such as burglary, vehicle crime, criminal damage and theft decreased in most areas during the evaluation period, while more spontaneous offences, such as violence against the person and public order offences did not. [43]

As presented in Chapter 1, the introduction of CCTV is associated with an increase of recorded violent crimes. Incidents that would have previously gone undocumented are now being observed by a CCTV operator and document. The literature suggests that CCTV must be used actively to have any effect on violent crime as passive effects of CCTV on violence were found to be insignificant [43].

2.2.2 Unfocused Human

The following quote by Virilio concisely conveys the benefits of using CCTV cameras when compared to human supervisors. In this subsection, I will present a review of the literature that more scientifically presents the underlying notion of Virilio's statement.

> Human Visual capability has difficulty competing with the high surveillance capabilities of the camera: the camera does not blink, sleep or get bored and, unlike images captured on videotape, the results of human visual surveillance cannot be rewound or replayed in a court of law. [140]

This quote concisely outlines some of the negative effects of being human in a surveillance position.To motivate the work presented in this thesis, the flaws associated with human-based operation of CCTV are discussed and used to justify a computer-based replacement or aide. Human flaws can be expressed as falling within one of the three categories, *vigilance*, *quantity* and *targeting information*. These categories all fall under the overarching theme of focus, a persons' ability to remain, with sufficient interest, on the task at hand.

Vigilance

An activity of interest can manifest itself at any time, and so cameras must constantly record footage to capture unforeseen events. It is imperative that a CCTV observer remains vigilant at all times so to identify any sudden events of interest. A study by Miller *et al.* [100] investigating U.S Coast Guard crew fatigue defines the concept of vigilance:

> Vigilance is the ability to sustain and focus attention in a boring situation, with the goal of quickly and accurately detecting the occurrence of a rare, unpredictable, important event. [100]

A report by Miller *et al.* [100] identified numerous factors that resulted in the decline of a person's level of vigilance. The factors, reported in order of importance, are age, hours slept, and hours worked. Similarly, a description of old, privately published research produced for the U.S Department of Energy was presented in a more recent paper by Green *et al.* [51]. The description of their previous research is as follows:

> Experiments were run at Sandia National Laboratories 20 years ago for the U.S. Department of Energy to test the effectiveness of an individual whose task was to sit in front of a video monitor(s) for several hours a day and watch for particular events. These studies demonstrated that such a task, even when assigned to a person who is dedicated and well-intentioned, will not support an effective security system. After only 20 minutes of watching and evaluating monitor screens, the attention of most individuals has degenerated to well below acceptable levels. Monitoring video screens is both boring and mesmerizing. There is no intellectually engaging stimuli, such as when watching a television program. This is particularly true if a staff member is asked to watch multiple monitors. [51]

Strong evidence for the hypothesis that human observation effectiveness decreases over time exists [51, 100, 109, 30]. Parasuraman *et al.* [109] presented a study investigating the ability of a human at detecting intent when handling a weapon. The authors found that image quality was an important factor when investigating the relationship between time and task effectiveness [109]. CCTV observation centres in the United Kingdom utilise legacy hardware that is known to produce poor quality footage. Considering the findings presented by Parasuraman *et al.* [109], one would expect to observe poor vigilance in some cases. A United Kingdom Home Office report reported that the amount of time an operator remains effective when given a continuous data stream is not conclusively known, but that the time after which effectiveness drops is somewhere between 30 and 120 minutes. When considering that a typical workday is 8 hours long, then it is expected that for the majority of a typical work day, a CCTV observer would be expected to under-perform for the majority of the time they are working.

Concisely, human operators are unable to maintain focus over long periods of time due to various factors. With this in mind, it is important to highlight that this imperfection in human visual attention longevity presents a foundation for computer-based assistance. Unless programmed to do so, a computer-based system does not suffer from longevity issues; they operate consistently as they cannot suffer fatigue, a major focal factor in the Miller *et al.* [100] study.

Quantity

It is important to note that a typical surveillance control room is fitted with numerous monitors that each show many live video feeds each. A study undertaken by Voorthuijsen *et al.* [139] saw several participants sit through two hours of recorded surveillance footage while noting any incidents depicted; the results showed that when presented with 4 video feeds at once, the detection rate drops from 91% down to 72% compared to monitoring a single video feed. The paper also notes that the performance may not be fully indicative of real life conditions as the test participants kept a high level of concentration throughout, this could be due to the short test time, a longer, more typical eight or twelve-hour test would better reflect real-life conditions. A similar study by Tickner *et al.* [133] found that when identifying suspicious incidents within and around a prison, the detection accuracy for 4, 9, and 16 monitors was 93%, 94%, and 64% respectively. Furthermore, they reported that suspicious incidents that took place further away from the camera were more likely to be missed. As a decrease in operator effectiveness is seen by introducing a few extra video feeds, what sort of effectiveness would be seen after adding an extra 8? or what about 67? or even 646? Gill *et al.* [43] explain that CCTV systems can be very large and so intuitively, large teams of observers would be required to effectively keep track of everything in view, this will however become cost ineffective after a certain point.

Target Information

Visual search tasks using CCTV can be aided by the use of high-level semantic target information that informs the observer of their target; this information can be determined by the observer and learnt over time by watching past footage of similar events [64, 65]. The identification of suspicious behaviour can be useful as a form of targeting information, as suspicious activity is often a precursor of crime. Howard *et al.* [65] demonstrated that the cognitive speed of scene evaluation for a visual search task locating *suspicious* movement was slow, which can have severe effects on active CCTV observation. The degree to which a person is acting in a suspicious manner is subjective. Norris *et al.* [105] observed surveillance system operators and monitored their behaviour alongside crime statistics.

> Categorical suspicion is the largest single type, accounting for one-third of all targeted surveillance. Thus the most frequent reason that an individual is targeted is not that of what they have done, but because of who they are, and operators identify them as belonging to a particular social category which is deemed to be indicative of criminal or troublesome behaviour. [105]

An observer applies less focus on information that they consider non-criminal in favour of data that they believe to be criminally associated. Personal experience is not the only informant of targeting information, as in one case, Norris and Armstrong [106] disclose that the practice of race-based targeting was encouraged. Both Norris and Armstrong, and Goold *et al.* [47] discovered that racial bias was present in CCTV observation, leading to one race being targeting disproportionally. If the information that guides an observers gaze, the targeting information, is derived from a place of trusted and verifiable intelligence then it can lead to a functional benefit as useless information is discarded. However, if the targeting is weak or incorrect given the task, then focus of an observer will be directed away from potentially important scenes.

2.3 Violent Behaviour Detection

In this section, a review of computer vision based methods of violent behaviour detection will be presented and summarised.

When in motion, violent behaviour is commonly associated with sudden increases in object velocity, acceleration. This association stems from the intent to injure or hurt. For one person to cause damage to another, they must transfer a substantial amount of energy (force) from their body to another. To cause damage, one must increase the rate of acceleration of a weapon or attacking appendage (Equation 2.1).

$$\text{Force} = \text{Mass} * \text{Acceleration} \tag{2.1}$$

Given this association, Datta [26] and Deniz [29] produced methods that detect violent behaviour by identifying high motion acceleration. Deniz notes that high acceleration often manifests itself as a *visual motion blur* which can be measured and used to describe the kinematics of a scene. The method proposed by Deniz operates by computing the low-pass filter that is required to transform one frame into the next frame within an image sequence. The low-pass filter represents the nature of the motion blur that has developed across frames as ellipses. The low pass filter C is computed using fast Fourier transformed representations of consecutive images (Equation 2.2), where $\mathfrak{F}(.)$ represents the Fourier transformation function.

$$C = \frac{\mathfrak{F}(I_i)}{\mathfrak{F}(I_{i-1})} \tag{2.2}$$

To detect ellipses, the low-pass filter C is transformed using a Radon transformation and projected using *vertical maximum projection* to produce a histogram in which peaks represent motion blur. The histogram is described using various statistics that are used to represent the kinematics of a scene. Visual motion blur is an artefact introduced by an image capturing device. Motion blur arises when the rate of image capture is slower than the rate of motion of an object; causing the capture device to record the visual composition of a moving object as it occupies different locations in a frame over time. The magnitude of the blur effect is often in proportion to the speed of the object, where faster motions are likely to induce a larger motion blur. The rate of motion blur is contingent on the capture device, meaning that if a camera with sufficiently high shutter speed is used to record violence, then no motion blur will be induced. Therefore, Deniz *et al.* [29] assumes that the recording device is insufficient for recording violence, but as technology progresses, this assumption may become invalid.

Many of the following methods within this discussion derive motion statistics from optical flow approximation. Optical flow approximation is the process of approximating the perceived motion vectors of objects between two adjacent images in time; the aim is to identify pixel or feature displacement and derive perceived motion. Optical flow can be described as the motion of pixel $I(x, y, t)$ that moves a distance of (dx, dy) over dt time. Many optical flow approximation methods assume that the pixel intensity does not change as it moves and time progresses, this is known as the brightness constancy assumption and results in Equation 2.3.

$$I(x, y, t) \approx I(x + dx, y + dy, t + dt) \tag{2.3}$$

From Equation 2.3, the Optical Flow Equation 2.4 can be derived. Optical flow approximation methods are defined by their approach at solving the optical flow equation [63, 87]. Naturally, methods that adhere to the brightness constancy assumption will perform poorly in scenes with rapidly changing illumination.

$$I_x u + I_y v + I_t = 0 \tag{2.4}$$

Where

$$I_x = \frac{\delta I}{\delta x}; I_y = \frac{\delta I}{\delta y}$$

$$u = \frac{dx}{dt}; v = \frac{dy}{dt}$$

Shifting focus back to violence, Datta *et al.* [26] implemented a structured approach to solve the task of measuring person-on-person violence by first determining a person's silhouette, and subsequently their head, for tracking. The third derivative of displacement (by time), known as *Jerk* (Equation 2.5), is then incorporated in the composition of the Acceleration Measure Vector (AMV) to describe violent motion.

$$j = \frac{da}{dt} = \frac{d^2v}{dt^2} = \frac{d^3x}{dt^3} \tag{2.5}$$

The authors justify the inclusion of jerk through data observation and state that "During violence, the motion trajectory of a person experiences a drastic change after being hit by the other person and jerk is an effective way to capture this behavior". In addition to AMV descriptor, the method of violence detection uses information about the orientation of limbs, "During violence, people raise arms and or legs, and hence the orientation of hand and or leg starts to change towards being parallel/negative to the ground plane". The method described by Datta *et al.* [26] assumes that a person's body is both visible and trackable. This assumption is breached when analysing CCTV footage within city centre environments due to occlusions caused by pedestrians in populated areas. Additionally, the limb orientation check assumes that the violence is captured using a camera with a side view of violence, which is not always true in naturalistic environments. The authors note that their system will malfunction when presented with either gang violence, or violence that involves wrestling.

Riberio *et al.* [113] introduce the Rotation Invariant Motion Coherence (RIMOC) feature representation which encodes the eigenvalues of second-order statistics extracted from a set of Histogram of Oriented Flows; the eigenvalue representation provides rotational invariance. The authors assume that the actions that comprise violent behaviour are unstructured, and therefore attempting to model *violence* that generalises to all instances of violence is difficult. Rather than use a binary classification scheme, a single class model is fitted using only data that represents non-violence. A codebook of features is generated using k-means clustering. This is followed by learning a set of codeword ensembles that encode the spatio-temporal structure between local features. A set of unseen features are then evaluated by determining the similarity between learnt and observed ensembles. If the dissimilarity between the observed and learnt feature ensembles is high, then the observed feature ensemble is considered violent.

The Scale-Invariant Feature Transform (SIFT) is an interest point detection and description scheme used to describe local regions within an image based on pixel intensity [86]. Nievas *et al.* [104] introduced a SIFT variant, Motion Scale-Invariant Feature Transform (MoSIFT), to detect and describe patterns of optical flow. Optical flow refers to the approximation of motion between two images in sequence across time, derived from correspondence in pixel intensity between images. The MoSIFT algorithm operates by performing SIFT interest point detection on appearance data and then removing any detected interest points whose position corresponds with a point in the optical flow fields with an insufficient optical flow magnitude. The MoSIFT descriptor is created using a SIFT descriptor concatenated with a similar encoding that is analogous to a histogram of optical flows. Nievas *et al.* [104] described spatiotemporal volumes using both SIFT and MoSIFT feature descriptors; The authors used ice hockey footage to demonstrate that their method could distinguish between scenes of one-on-one fighting and standard play. Xu *et al.* [154] used the MoSIFT framework and demonstrated that classification could be improved through the use of a KDE based feature selection and a sparse encoding scheme.The KDE feature selection was used to remove irrelevant and redundant information by computing the probability density function of a MoSIFT descriptor; out of 256 features, only 150 with the greatest value in the probability density function are retained. The reduced feature representation combined with sparse coding results in an increase in classification performance when compared to the defaul MoSIFT algorithm.

Gracia *et al.* [49] present a real-time violence detection system that operates using computationally cheap methods of image analysis as they argue that approaches such as MoSIFT, although impressive, are too computationally costly to be practically implemented in the real world. Their proposed method operates by performing adjacent frame differencing and applying a fixed threshold to generate a binary representation of objects in motion. Blobs are extracted from the binary representation, and the largest blobs are then described using measures of inter-blob distance and compactness. This method does not provide state-of-the-art performance but does provide reasonable performance at a substantially reduced computational cost. The authors note that their method has trouble analysing slow, continuous motions as differencing between frames will result in small blobs. This method assumes that violent actions cause the largest displacement of objects over time as implied by the author's decision to describe a scene using the *largest blobs*. This assumption may fail when considering the nature of CCTV footage in city centre locations. CCTV often depicts people at various depth levels relevant to their distance from the camera. Therefore actions occurring closest to the camera will likely generate the largest blobs, irrelevant to the action.

Fisher Vector encoding represents a visual scene by spatially pooling local features. Bilinski *et al.* [5] introduce a spatio-temporal Improved Fisher Vector (IFV) representation and compares performance with the traditional IFV approach using violent behaviour detection datasets. The authors describe a computationally efficient implementation for fast encoding of spatio-temporal features. The authors report high classification accuracy on violence detection tasks, however there is a caveat concerning computational cost . Although the proposed feature encoding can be performed in real-time, the utilised local feature extraction method (Improved Dense Trajectories [143]) can not. Computationally cheaper feature extraction would be necessary for a system that utilised their proposed encoding scheme to be deployed for use in active surveillance scenarios.

Hassner *et al.* [56] introduced the Violent Flows (ViF) dataset containing instances of violent and non-violent crowd behaviour, typically extracted from footage of crowds attending sporting events. Alongside the dataset, the authors proposed an optical flow based descriptor. The violent flows descriptor is formed by first computing the absolute difference in optical flow magnitude between two successive frames. A threshold is applied to create a binary representation where 1 indicates that the change in flow magnitude is greater than the average change in flow magnitude, otherwise 0. A temporal set of binary maps that represent points with strong changes in optical flow are averaged temporally before being spatially pooled and used to populate a fixed-size histogram. The Violent Flows representation encodes patterns of relative change in optical flow magnitude over time. Gao *et al.* [40] suggest that encoding changes in motion orientation is important when analysing violent crowds and suggested the ViF variant, Oriented Violent Flows (OViF). As an independent feature, OViF does not outperform ViF at classifying violent data, but when used in combination with ViF, provides complementary information that increases performance above ViF levels.

Brémond *et al.* [13] implement an ontological approach for behaviour detection in European metro stations. Using tracking systems, the authors monitored individuals, groups and crowds and utilised an ontology to determine the perceived behaviour. The ontology, defined by Brémond *et al.* [12], is used to determine whether a camera captures footage of vandalism, overcrowding, persons jumping over barriers, and persons blocking someone's path, and fighting (violence). The ontology defines four rules indicators of violence, if one of these is satisfied, then violence is detected. Violence is detected if a person is lying down, if group width variation changes significantly, if a group splits quickly, or if there is significant variation in motion trajectory amongst the group. The ontology is applied to a constrained environment in which typical behaviours are clearly identified. It is arguable the same ontology would be unsuitable for the NTE as an NTE environment is both physically and socially different compared to a metro station. For instance, alcohol consumption affects gait [28, 68, 132, 110], leading to unstable movement that may break ontology rules associated with group width variation and group splitting without actually being violent.

Marsden *et al.* [93] present a set of *holistic* features that represent attributes of crowd motion, these being *collectiveness*, *conflict*, *density* and *mean motion speed*. These features are derived from Kanade–Lucas–Tomasi (KLT) tracked motion trajectories [134]. *Collectiveness* describes crowd movement uniformity, and *conflict* encodes interaction between neighbouring trajectories. Whether violence exhibits greater *conflict* or *collectiveness* than non-violence is not explored. More recent research by Marsden *et al.* [94] investigated the multi-objective training of deep residual networks. It was reported that training a deep network to simultaneously solve *crowd counting*, *violent behaviour detection* and *crowd density estimation* resulted in a model that can detect violence to a greater degree than a model trained exclusively to solve the *violent behaviour detection* task. This suggests that more informative representations of crowd dynamics can be learnt given alternative crowd based objectives.

Meng *et al.* [99] used a pre-trained VGG-19 convolutional neural network to extract a *spatial* and *temporal* stream to encode patterns that occur in pixel intensity and short-term motion respectively. The authors demonstrate that the detection of violent behaviour is greatest when utilising both appearance and motion information. Furthermore, through the addition of Improved Dense Trajectories (IDT) [143], the authors demonstrated that long-term temporal motion is beneficial when differentiating between violent and non-violent samples. Dong *et al.* [32] utilise Long Short Term Memory (LSTM) networks to extract *spatial*, *temporal* and *acceleration* streams for the task of violence detection. An LSTM encodes variable length sequential patterns by selectively storing important information that defines an action. This research demonstrated that acceleration provided the greatest performance for one-on-one violence detection when considered independently from *spatial* and *temporal* streams. The combination of all three streams provided the best performance. It is important to note that each method that concerns *deep learning* utilised pre-trained models and that any model fitting was performed using fine-tuning, updating an already trained model using new data. In Chapter 5, Section 5.7, it is demonstrated that training from scratch results in a poor model of violence, likely caused by the small amount of available data.

2.3.1 Conclusion: Detecting Violent Behaviour using Computer Vision

To conclude, the ability of a human completing a video-based task of action detection using footage captured by CCTV is typically poor. The factors associated with *vigilance* and *quantity* have been experimentally studied whereas issues associated with observer bias has been anecdotally observed. Human observation ability is sub-optimal given a large amount of data, and often scenes of interest go unobserved. The mass adoption of CCTV in the United Kingdom has created a potential avenue for improving the negative effects of violence. Increased CCTV allows for the capture of incidents that would previously go unreported and unobserved.

Identifying scenes of violent behaviour allows for appropriate action to be taken in an effort to reduce the severity of an injury and the associated cost of treatment (see Chapter 1). Computer vision methods for violent behaviour detection presented in the literature typically focus on achieving highly accurate systems, often with an associated high computational cost. In order to be applicable in an active surveillance situation, algorithms are required to operate in real-time to handle the constant stream of data sent to CCTV observation environments. Use of a real-time vision system is highly important when attempting to reduce the severity of the sustained injury as intervention time is the key factor. An algorithm that imposes a substantial delay between the time of an action being performed and time of detection will fail to fulfil the criteria associated with the research hypothesis. Concisely, computer vision can improve human ability at detecting violent behaviour which has a theoretical positive influence on the severity of the sustained injury.

Research using real-world CCTV footage is uncommon, especially regarding NTE hours in the UK, which are characterised by their increased levels of violence. Frequently, the Hockey violence dataset is used to evaluate violent behaviour detection techniques [5, 29, 49, 56, 104, 154], however there is no evidence to suggest that Hockey violence provides a suitable analogue for real-world street violence. Similarly, research by Riberio *et al.* [113] and Brémond *et al.* [13] evaluate their proposed methods using unique datasets that each have a specific context; violence and aggression in train/metro station environments. Train/metro station datasets have a specific context that are not guaranteed to be representative of violence in city centre NTE environments. Lastly, datasets like BEHAVE [8] or PETS2015 [79] use actors to portray scenes of violence. Acted scenes may not provide an accurate representation of violent behaviours seen in real-world NTE videos as actors will behave in a manner such to prevent inflicting pain. Using this data to generate a detection model may lead to a model that generalises poorly when applied to real-world data. To summarise, research is frequently conducted using datasets that do not necessarily represent natural scenes within UK streets. Analysis should be conducted using real-world data to allow for a genuine understanding of how violence detection system will perform in real-world situations.

2.4 Predictive Modelling

So far, the discussion presented has concerned itself with the development of computer vision based models for the *detection* of violent behaviour.

The methods described in Chapters 4 & 5 can be applied as reactive systems, providing information that allows law enforcement to react in an informed manner as an event unfolds. The *prediction* of violent behaviour would allow for preventive measures to be deployed. Unfortunately, the video data collected (Chapter 3) does not include information that displays the evolution from non-violence to violence. Therefore, generating and evaluating predictive models using video data and the computer vision methods is not feasible.

During data collection, Global Positioning System (GPS) location data for violent incidents in Cardiff and Northampton was obtained. Using this information, models that predict violent crime hotspots revealed by the GPS data are generated. Agent-based modelling provides a simulation of pedestrian movement dynamics. Within this section, a review of the literature associated with agent-based modelling is presented. Chapter 6 will present implementation details of an agent-based model aimed at simulating city centre environments. An evaluation of the ability of an agent-based model at predicting violent crime is presented.

In the discussion and closing remarks to this thesis (Chapter 7), a theoretical foundation disclosing how systems of detection (Chapters 4 & 5) can be integrated with systems of prediction (Chapter 6) will be presented. This will form the basis of future work.

2.4.1 Agent Based Modelling

Criminal behaviour is believed to be influenced to some degree by the environment [131], resulting in a concept in the field of crime prevention which asserts that non-linear distribution of police resources can minimize the effects of crime. This concept stems from the idea that certain spatial areas in a town or city require less police presence than others. By varying the level of police patrol units and monitoring subsequent changes in crime, it was determined that the amount of allocated police resources had no significant influence on crime [71]. This research provides contradictory evidence to the assertion that non-linear distribution of police resource and maximize crime reduction. Sherman *et al.* [123] claim that the aforementioned research suffers methodological errors, and in response, developed their own study on the same topic. Sherman *et al.* [123] presented a study which demonstrated that non-linear distribution of police resources was beneficial. In their study, they identified 110 hotspots using historical crime data. Fifty-five hotspots were given increased police attention whereas the other fifty-five were used as a control. The study purports a reduction of 13% in reported crimes in the areas with increased policing, suggesting that non-linear distribution of police resources does have an effect on crime. More precise crime mapping allows you to understand whether a specific establishment along a street is a contributor to a certain type of crime. Multiple studies have demonstrated that the rate of crime can be reduced across many types of criminal activity, such as robbery [123] and different forms of violence (gun [122], physical [10]). However, the efficiency of hotspot policing is dependant on the type of crime [35]. Simply increasing police presence does not yield the maximum reduction in criminal behaviour possible, benefits in crime reduction can be improved by identifying the core factors that are associated with a crime hotspot and applying a more bespoke policing strategy [9, 33]. The benefits of hotspot analysis are not tied to the hotspots themselves as:

> The results of our research suggests that hot spots policing generates small but noteworthy crime reductions, and these crime control benefits diffuse into areas immediately surrounding targeted crime hot spots. [9]

Police currently analyse historical data to determine hotspots. This form of data analysis does not allow data analysts to understand how the distribution of crime changes when presented with a different environment than the one crime initially took place within. Crime hotspot predictions can be generated using factors identified as being related to criminal behaviour. This has applications in city planning and management; in the context of the NTE, a model that can predict the nature of violent crime would allow the local governing body to best determine how to design the night time economy environment so to minimize the effects of violence.

Pedestrian Modelling

Work presented in Chapter 6 proposes the simulation of pedestrian behaviour within NTE environments with the aim of determining regions of associated with violent crime. Computer simulations have been developed to mimic the movement of pedestrians so as to offer a better understanding of the effects of environmental design on the behaviour of moving pedestrians. Prominent work presented by Helbing *et al.* [60] introduced a social force model. In this model, independent simulated agents were guided by rules governing their motivations and actions. An agent has a desired spatial goal which they will move towards using a propulsion force, synonymous with the desired velocity of that agent. A repulsion force is applied to avoid collisions with the environment and other agents. Using basic rules applied to each agent in a multi-agent model, emergent behaviours representative of those exhibited by real-world pedestrians arose. For instance, in crowds of oppositely moving directions, lanes would naturally form as they allow agents with a common course of motion to proceed towards their goal more efficiently. The environment affects the emergence of behaviours such as lane formation. In situations where pedestrians walk in opposite directions, appropriately placing obstacles more easily induces lane formation [61]. Environments can be computationally designed and tested using simulations to alter pedestrian behaviour [61, 157, 158].

Agent Based Modelling (ABM) has been used to aid risk assessment and disaster planning; by simulating behaviour during potentially harmful events, responsive measures that reduce injury and prevent loss of life can be determined. Comparing route planning strategies for the task of evacuating a metro station, Zhang *et al.* [158] demonstrated a reduction in evacuation time when simulated agents considered the local density of other agents when selecting an evacuation route. Similarly, SAFEgress, and ABM framework, has been used to investigate how signage and local situational information affects routing choices during disaster scenarios [19]. Further evacuation-based examples include the investigation of exit strategies across various building types [153] and the production of risk maps [111] in response to earthquakes. By combining an ABM with a hydrodynamic simulation, Dawson *et al.* [27] produced a dynamic ABM that simulated pedestrian movement during a flood event. The flood simulation is used to estimate the vulnerability of individuals to flooding under various conditions and test flood prevention measures. Research presented so far has been concerned with evaluating pedestrian risk given certain circumstances, actions and events. Rather than assess risk from a pedestrian perspective, Hawe *et al.* [57] used risk to inform an ABM for resource management. A simulation was developed to identify the most effective allocation of emergency resources such that priority is given to the most critically injured pedestrians in a disaster scenario. In the context of crime, ABM can be used to test new police strategies as shown by Dray *et al.* [33] whose model reported that problem oriented policing is better than random patrol, which is accurate to reality [9]. Numerous factors are assumed to influence harm in the NTE, including premises opening hours, congestion, population, blood alcohol level, etc. The authors of SimDrink present a proof-of-concept model that predicts *harms* (violence) during a typical night out in Melbourne [119]. The SimDrink model provides and experimental platform that is intended to provide an understanding on how *harms* are affected by policy changes such as altering venue operating hours.

There are different types of models for the simulation of pedestrian behaviour. The type of model used in the experiments described in this thesis is a model that provides continuous spatial freedom for each simulated agent, this model is known as an Agent Based Model. An alternative model that arises in the literature is one that sees the positions of each simulated agent being tied to a cell within a pre-defined spatial structure. In the case of Cellular Automata, physical environments are represented using a matrix structure, where each element represents an occupy-able space by some abstract entity, in the case of this work these are pedestrians. The state of each cell in the matrix updates at each simulated discrete time-step based on the state of neighbouring cells. Due to the spatial constraints imposed, a Cellular Automata (CA) model is typically computationally less complex when compared to an Agent Based Model due to the finite number of possible interactions between simulated agents [7]. However, CA models have been demonstrated to give realistic emergent phenomena (lane formation, side-stepping, herding, crowding, etc.) during evacuation and normal scenarios [159]. The work presented in Chapter 6 builds upon an existing Agent Based Model that models drunken behaviour [101].

2.4.2 Alcohol, Anger and Stress

Moore *et al.* [101] have previously described the theoretical mechanisms relating alcohol intoxication to pedestrian movement and outcomes that include aggression and violence. Briefly, undesired events that people have no control over trigger negative emotional responses including stress and anger [20, 114]. This can arise in cases where many pedestrians occupy a small space as competition for free space can elicit anger [67]. Furthermore, pedestrians in crowds will experience repeated invasions to personal space, determined as being within a 1.2-metre radius of another pedestrian [52]; this also elicits stress [128]. Stress caused by the invasion of personal space induced is moderated by familiarity with the person invading personal space and familiarity with the environment itself [52, 58, 69] and culture [62]. Violence is more likely under conditions of stress [22, 73, 95, 103]. For example, stress is positively associated with hostility in men towards women [4] and being stressed makes it harder to cope with frustration [95].

Based on the literature, it can be argued that violence and aggression in crowds are partly attributable to pedestrian congestion eliciting goal inhibition and stress. For the NTE, however, the additional condition where a proportion of pedestrians are intoxicated to the extent that their gait becomes unstable will contribute to this. Intoxication effects more than gait, the perceptual [14] and mental of pedestrians are known to change. Adaptive emergent behaviours, like lane formation and queuing, typically improves pedestrian flow. These and other adaptive behaviours may be disrupted in the presence of intoxication as pedestrian gait is affected. Experiments are used to assess whether including variable levels of intoxication in an ABM, realised as modifying pedestrian gait, provides additional explanatory power in simulations designed to identify the likelihood of violence in NTE crowds.

2.4.3 Conclusion

Understanding where and when disorder and violence will occur is of value to efforts aimed at mitigating harm: this can be achieved by altering physical or social environments through means such as pedestrianising streets, or by adjusting the opening times of pubs, clubs, and bars. Agent-based models provide in silico simulations of real-world pedestrian flow. In many night-time locations, alcohol is used as a social lubricant and is synonymous with an unsteady gait and violence. The significant value of ABMs is that they can be used in a variety of contexts to provide risk assessments and inform environment design to minimise harm. ABMs have not been rigorously applied to NTE contexts, and it is not clear whether simulations can provide insights to inform the design of the NTE.

2.5 Summary

A review of the literature presented in Chapters 1 and 2 suggests the sub-optimal use of technology for the task of violent behaviour analysis. More specifically, the literature presented in this chapter highlights two key areas of research, computer vision and pedestrian simulation.

Regarding video surveillance, identifying scenes of violent behaviour can result in a reduction in the severity of an injury and the associated cost of treatment (see Chapter 1). Reviewing existing methods of violent behaviour detection while also considering the research hypothesis reveals that current methods are not sufficient as real-time performance is rarely a concern. Given that injury severity is contingent upon intervention time, an algorithm that operates with a significant temporal delay would not satisfy the research hypothesis. Across Chapters 4 and 5, two methods of violent behaviour detection that both operate in real-time while each maintaining a high detection accuracy, are presented.

Reviewing the literature on pedestrian simulation and ABM revealed that the effects of drunken characteristics on emergent behaviour and crime prediction have not been widely investigated. The value of ABM is that it can be used in a variety of contexts to provide risk assessments and inform environment design to minimise harm. This type of model can be beneficial when you consider that the risk of engaging in a violent altercation increases during NTE hours. Presented in Chapter 6 is an investigation into the efficacy of alcohol informed pedestrian simulation model for the task of violent behaviour detection.

Chapter 3

Video Data Collection and Data Bias Analysis

3.1 Introduction

Chapters 4 and 5 of this thesis investigates the development and application of computer vision techniques for the detection and analysis of violent behaviour captured using CCTV. This chapter will provide an overview of the data collection procedures performed to gather video data for analysis. In summary, our data collection used two sources; online repositories used within existing research and privately sourced data from police organisations in the UK.

Presented in this chapter is an overview of six datasets alongside an investigation into the biases that exist within each dataset. Data bias arises from inconsistencies in the *data capture*, or *data annotation* process [135]. As an extreme example of data capture bias, a violent behaviour dataset may contain instances of violence recorded during night time hours, and instances of non-violence recorded during the day time hours; the difference in lighting condition induces a bias, leading to a model that associates darkness with violence. In this example, the model will fail to detect violence that occurs during the day.

Highlighting data bias will help understand the suitability of different datasets for the task of violent behaviour detection and analysis. Classification results reported using a dataset whose biases are prominent and easily exploited should be investigated thoroughly to avoid misrepresentation. In this chapter, it is demonstrated that state-of-the-art performance for detecting violent behaviour can be achieved by exploiting data bias; information that has no theoretical relationship with violence can be used to identify violence. This information is useful for informing computer vision system design.

3.2 Data Collection

Work presented in Chapters 4 and 5 describe the utilisation of video data for the generation of models for detecting violent behaviour. The first and foremost task to complete was data collection, as without data, we lack an essential component for performing data analysis. Data was gathered using two distinct approaches. The first was to undertake a literature survey and identify publicly available datasets suited towards my thesis subject matter. The second form of data collection involved obtaining private data from police organisations, who are required to store video data for various legal purposes. Before my induction into the PhD program at Cardiff University, Prof Simon C. Moore had arranged an information sharing agreement with South Wales police, which allowed access to a small amount of data captured from CCTV cameras within and around the Cardiff area. Unfortunately, gathering more data from this source was not possible due to staffing issues on their end. During early research, I attended a government-led conference on data access for crime prevention known as Video Analytics for Law Enforcement or VALE for short. The objective of VALE is to identify the areas of law enforcement which will benefit from the introduction of computer-based video analytic techniques. A key point of discussion at the conference was the lack of data centrality; data was stored and managed by each police organisation independent of any other. A goal of collating all the video information into one extensive collection and providing a central point of access was proposed. As stated at the conference, the VALE program was not official at the time of discussion, and its fate was subject to upcoming political events. Information on VALE disappeared shortly after, which was likely due to political turmoil during the general election at the time. Although this point of data access is no longer available, it did highlight that directly contacting the local police organisations may allow us to secure data. A data sharing agreement between Cardiff University and Northamptonshire Police was developed to allow for data access. The agreement allowed for the collection of both violent behaviours captured on CCTV and GPS data related to all crimes within the Northamptonshire region (see Appendix A).

3.2.1 Northamptonshire Data Agreement

A data sharing agreement was established between those involved with this thesis and Northamptonshire police. In this section, I will outline the challenges associated with collecting and processing video data included in the agreement. Although I was vetted (Non Police Personnel Vetting (NPPV) Level 2) as per the agreement, I was not allowed direct access to the criminal record storage room located in Northampton. The process of obtaining data required that I manually identify records of interest using a computer system and ask a particular member of staff to retrieve the physical documents. It was typical for the CCTV disc in the physical record folder to be either broken or missing, making the data collection process tedious. The failure to preserve data may be a breach of the Data Protection Act that may have wider ramifications if any case with missing evidence is recalled to court. During collection, I recovered all records marked as "Violent Behaviour" between the years of 2013 - 2015; everything prior to 2013 is stored in a long-term archive building that requires a different process to access.

The video data collected was stored in a proprietary format designed by CCTV security firm PELCO. A video player was provided but did not offer a means of making the data processable by our algorithms. My initial attempt at extracting the data was to play the video files using the propriety player and directly record the screen. This approach did not work as the video viewer displayed videos at an inconsistent frame-rate resulting in the screen recorder either duplicating or missing frames entirely.

Analysing the proprietary video format using BinWalk, software used to identify common structures in data, it was determined that the video frames were stored in JPEG format. The JPEG images could be separated into two groups. The first group were images that had the same dimensions as the final video as determined by the video viewer. These images could be considered key-frames as they were visually complete and contained no artefacts bar those expected from video compression. The second group of images had dimensions of $16 \times N$ where N was a seemingly random number being a multiple of 16.A set of pairwise values corresponding with (x, y) co-ordinates where $0 \leq x < videowidth/16$ and $0 \leq y < videoheight/16$ were identified by analysing the hexadecimal data within the video file.It was hypothesised that the proprietary format used a Motion JPEG type of encoding where key-frames were stored and only fixed sized blocks that change over time are saved, 16×16 image segments were extracted from non-complete images and placed onto the last complete image using the (x, y) positional values.

3.3 Crime and Intoxication Data

GPS crime data and intoxicated pedestrian simulation are used to create a predictive model of violent crime (Chapter 6). GPS crime data was obtained from both Northampton and Wales Police forces. The information included with the GPS reports provided by the Police described *time*, *longitude*, *latitude* and *crime type*. There are various types of crime recorded by the Police, but for the work presented in this document, only crimes related to violent behaviour were used. The specific category of crime used was Violence Against Person (VAP). The accuracy of GPS crime coordinates from Northampton are quoted to have up to a 1.5-metre deviation from the true location. GPS error was not provided for the South Wales Police data.

Regarding intoxicated behaviour modelling, Blood Alcohol Concentration (BAC) data gathered from the city of Cardiff was obtained using data sources associated with existing research [110]. The BAC dataset includes the level of intoxication (BAC) of pedestrians within Cardiff during the NTE. The BAC dataset also contains a categorical variable indicating the observed state of a persons gait as being either *normal* or *staggered*.

3.4 Video Datasets

Six video datasets are used throughout this thesis; four of which are datasets dedicated exclusively for video based violence analysis. Two of these have been sourced by Cardiff University using data sharing agreements; both South Wales Police and Northamptonshire Police agreed to share their data CCTV data. Two crowd abnormality datasets have been used to determine how well the violent crowd detection method (Chapter 4) generalizes for detecting non-violent scenarios that are likely to occur in city centre environments. The six datasets used throughout this thesis are:

Cardiff Dataset: The Cardiff Violence Dataset (CF-Violence) was sourced from South Wales Police. Data collection was performed by a member of their staff. The dataset provided contained approximately 23 hours of footage, of which 1 minute 56 seconds was of violent behaviour. To increase the amount of data, an extra 90 seconds of violent footage was downloaded from online video repositories and added to the dataset. Footage was selected if it was manually identified as being visually similar to data sourced from the police. (Figure 3.1). In total, the Cardiff dataset contains 13 scenes of violent behaviour. In addition to normal behaviour, the data provided contains two scenes of non-violent aggressive behaviour and 3 instances of post-violence that depict the victim of violence being tended by medical personnel. Due to unbalanced nature of the dataset, a random subset of normal behaviour equal in size to that of the violent data is used during training.

Northampton Dataset: The Northampton dataset (NN-Violence) consists of 18 police records from 18 unique cases, providing 18 instances of violence with a varying amount of normal data captured before and after the incident. Each police record contains footage captured from multiple cameras providing different view points for some of the violent incidents. Even though a single case record has multiple camera views of an incident, only the best view of an incident is stored at 30 frames per second at the maximum resolution the capture device is capable of (typically 704 × 576); all other views are stored at 4-6 frames per second at a resolution of 704 × 288.

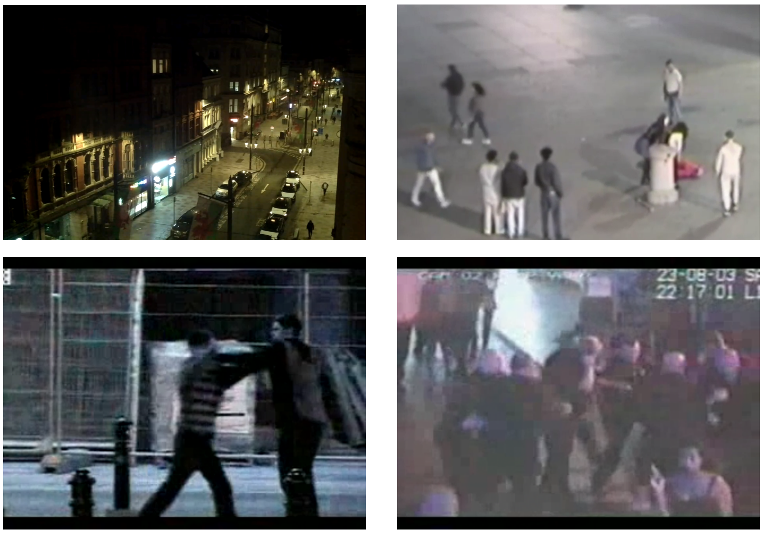


Figure 3.1: Still images identified as being visually similar to data found in the CF-Violence and NN-Violence datasets.

Violent Flows: Introduced by Hassner *et al.* [56], this dataset consists of 123 violent and 123 non-violent videos of crowds within and around sporting stadiums. As described by Hassner, the data was extracted from Youtube and therefore contains heavy compression artefacts. This dataset is applicable for this thesis as sporting areas are typically built within city centre environments (Figure 3.2).

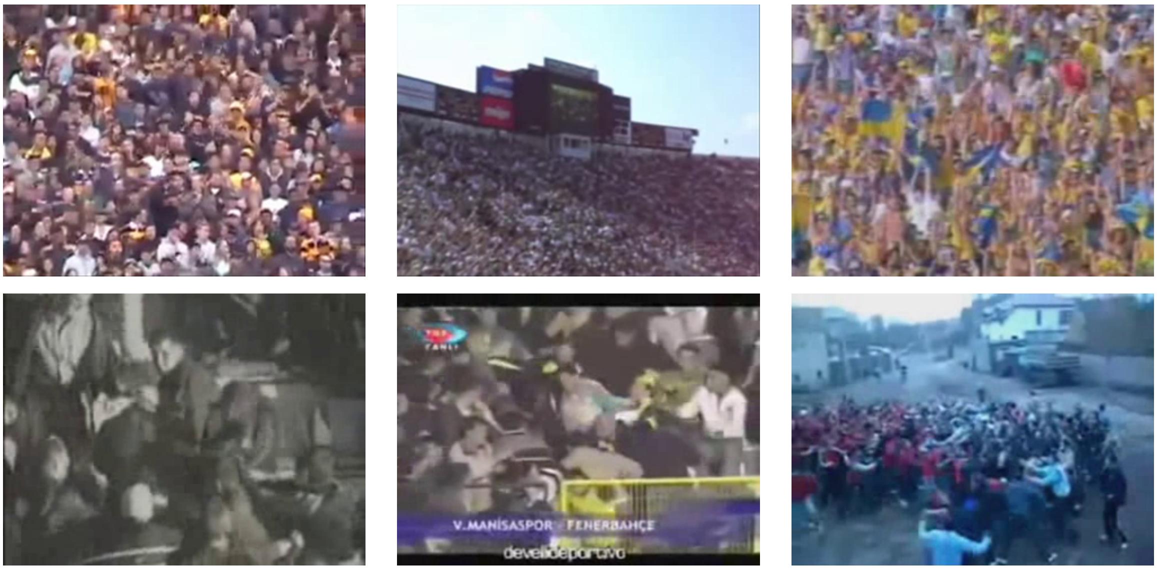


Figure 3.2: Still images extracted from the Violent Flows dataset.

Hockey Violence: Introduced by Nievas *et al.* [104], the Hockey Violence dataset contains 500 samples of non-violent, and 500 samples of violent scenes captured from televised Ice Hockey matches (Figure 3.3). It was observed that the visual composition of violent scenes found within this dataset is similar to one-on-one brawls that are typically found in city centre environments. This observation has yet to formally backed up using quantitative analysis that compares street violence and hockey violence.

UCF Web Abnormality: Introduced by Mehran [98], the UCF dataset contains 20 samples of video footage depicting crowded city environments during both normal and abnormal behaviour. The 8 videos of abnormal behaviour

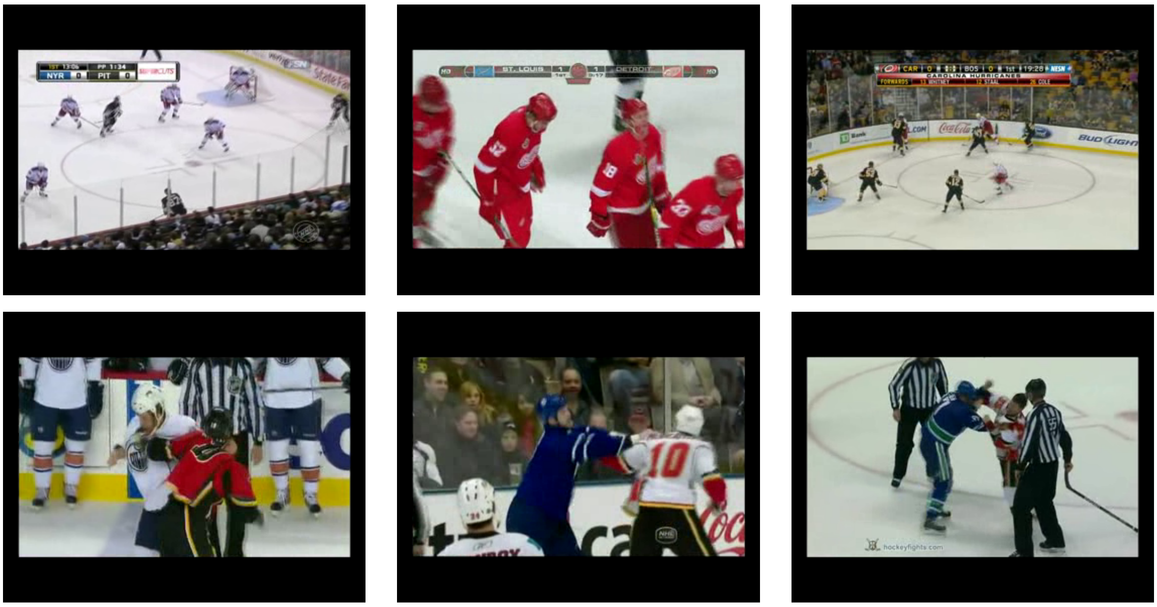


Figure 3.3: Still images extracted from the Hockey violence dataset.

within the UCF dataset each falls into one of three sub-classes; panic, clash and fight. Footage is recorded using static cameras; each scenario is subject to changes in illumination and non-pedestrian objects such as vehicles and flags. The range of normal behaviours include walking, jogging, queueing and timed road crossings (Figure 3.4).

UMN Crowd Panic: Introduced by Mehran [98], the UMN panic dataset contains footage of a physically simulated panic type scenario. A group of up to 14 people is instructed to partake in standard social behaviour until being told to flee in unison so to simulate a large disruptive event. The footage is captured using static cameras pointed towards a controlled environment where no external influences are present with the exception of very minor global changes in illumination (Figure 3.5).

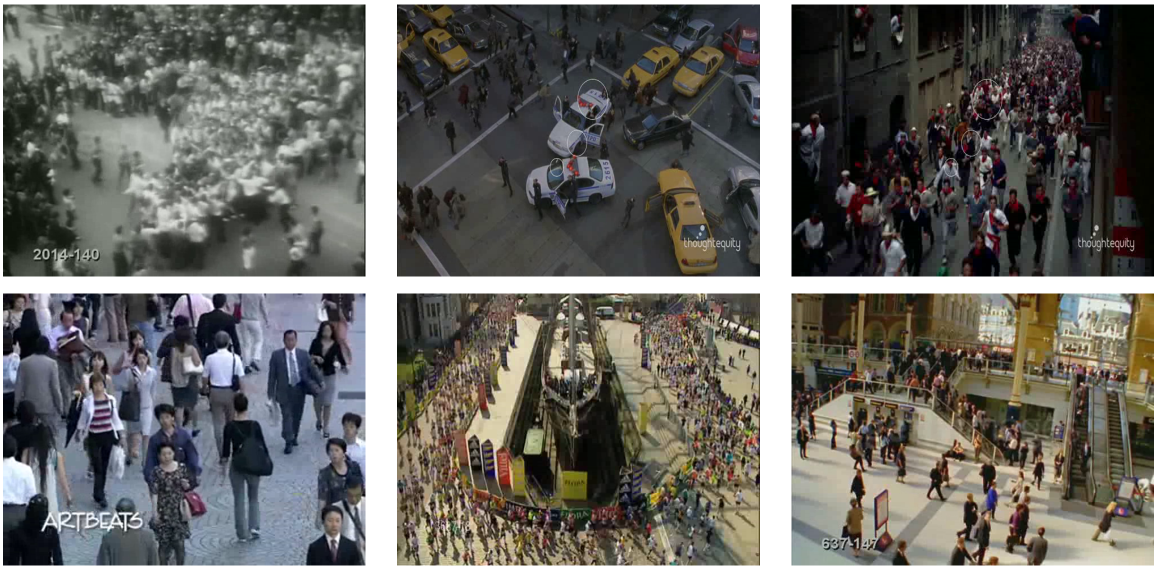


Figure 3.4: Still images extracted from the UCF Web Abnormality dataset.

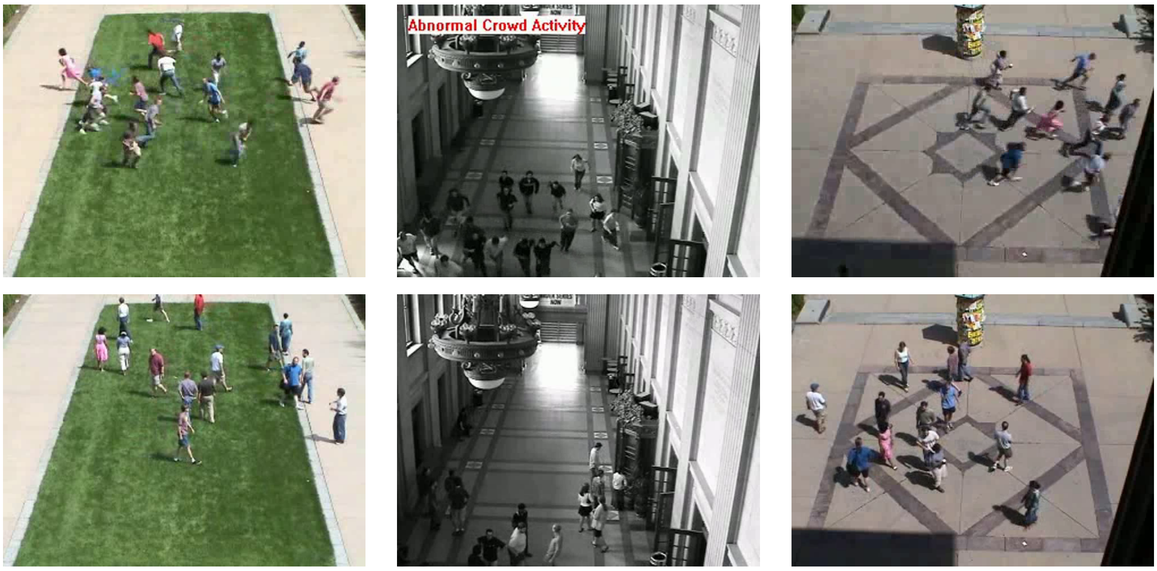


Figure 3.5: Still images extracted from the UMN dataset.

3.5 Investigating Bias: Quality and Depth

Biases in the data may be introduced during data capture. Systems with different image capture characteristics may, by accident, introduce a bias in which scenes of abnormality (violence) are represented differently that scenes of normality. In this section, an analysis of perceived image quality and depth of footage contained in each video dataset is presented. The hypothesis that aspects associated with image capture characteristics can be used to classify between scenes that depict abnormal (violent) and normal behaviour is tested and discussed. The purpose of this analysis is to demonstrate the internal biases of each dataset that may lead to computer vision based techniques to achieve state-of-the-art classification performance by describing aspects truly unrelated to the action of abnormal behaviour. In the conclusion, each dataset is ranked based on their data capture biases, a ranking that should be considered when discussing the ability of a method to detect abnormal (violent) behaviour correctly. Finally, suggestions for avoiding certain types of biases will be presented alongside a discussion for potential future work.

3.5.1 Image Quality Analysis

The perceived quality of an image is affected by the image capture process. Regarding digital media, both the hardware and software can introduce various distortions or artefacts during image capture. In this section, an investigation into the perceived quality of video images within each dataset is presented. Scenes of abnormal (violent) behaviour may be recorded under different conditions than scenes of normal behaviour, for instance, it was observed that footage of violence in the CF-Violence dataset is often zoomed in (demonstrated in Section 3.5.2); the act of applying a digital or analogue zoom may introduce noise or distortion. By applying Image Quality Assessment (IQA), image quality characteristics can be measured and then use to determine whether an appearance based bias is present in the data. A bias in perceived image quality may allow for effective discrimination between violent/non-violent behaviour based on information that is unrelated to the violent or non-violent behaviour of captured persons. A model that exploits image capture characteristics (quality) for violent behaviour detection may not generalise well as data recorded using another image capturing system whose capture characteristics are different, will be poorly classified. Presented in the following section is an evaluation of the image quality properties present in each dataset. Information and knowledge derived from IQA can be used to inform feature design and aid in debugging of unsupervised machine learning techniques.

Two branches of IQA are used for evaluating image quality, these are *no-reference* (blind) and *full-reference* assessment. *Full-reference* methods of IQA utilise a reference image for comparison with an image that suffers from degraded image quality; a measurement of the differences between images is used to quantify perceived quality [120, 148]. *No-reference* methods apply models that extract and measure properties of image quality [2, 55, 96]. Research has shown that perceptual interpretations of quality are typically subjective, and therefore many models of IQA are based on human perception [2, 55, 89, 96]. In the following analysis, an overview of each IQA method is presented alongside classification scores achieved by applying 5-fold cross validation for the task of classifying between normal and abnormal (violent) scenes. In addition to classification results, distribution plots for each each IQA measure for both normal and abnormal (violent) samples are included. To make statements about the typical composition of abnormal (violent) scenes, the difference in measurement distribution between behavioural classes is statistically tested.

The testing methodology for this section is first to present an explanation of each metric of image quality alongside a description of the results with relation to each dataset; a discussion and conclusion will follow this.

No-Reference (Blind) Image Quality Analysis

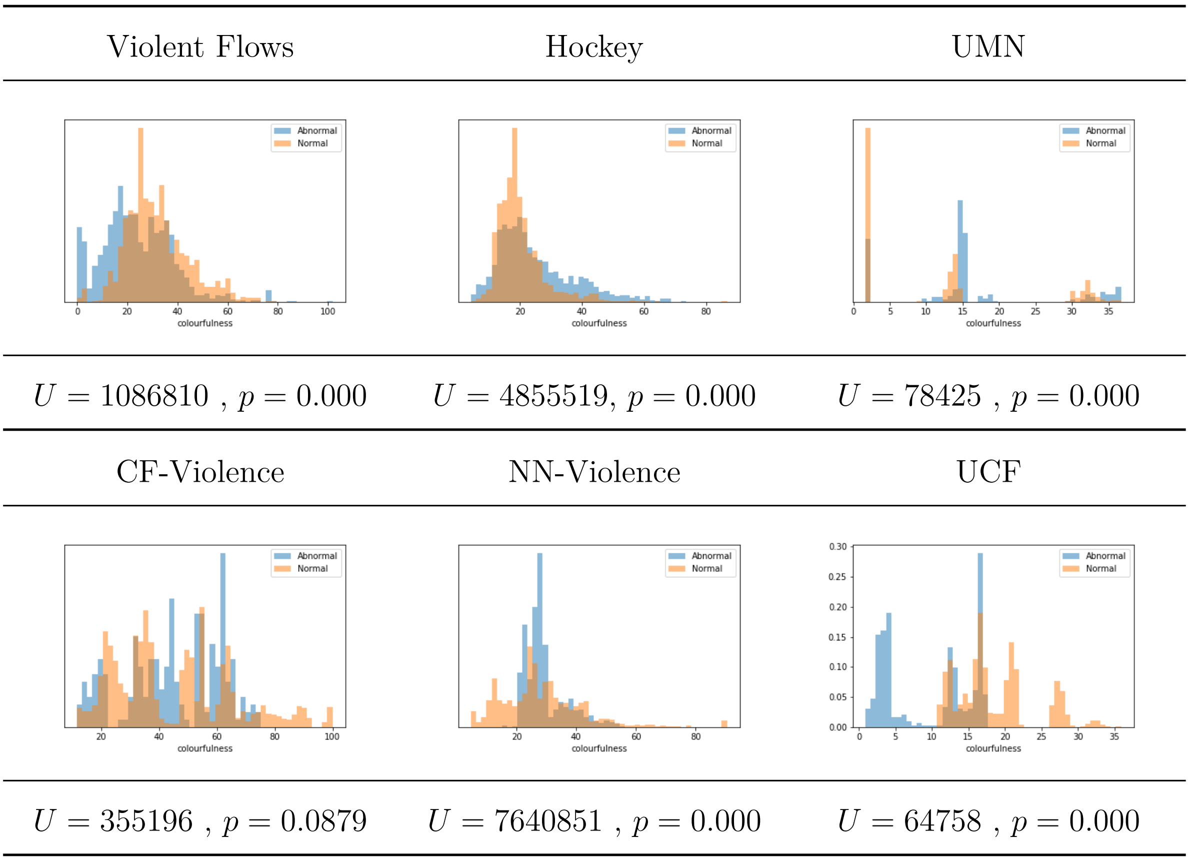
In the proceeding subsections, multiple methods of no-reference IQA from the literature are introduced. The characteristics of no-reference image quality that are investigated are: *Colourfulness*, *Sharpness*, *Contrast*, *Complexity* and *Noise*. These characteristics were selected as they are well defined and widely researched properties of image quality. For each property, the distribution of values populated by measurements of a particular image property computed using each frame of within a video for a given dataset is reported. Statistical significance testing methodologies are applied to determine whether the difference in image quality characteristics between violent and non-violent samples is significant. Additionally, classification is performed to demonstrate to what degree a violent action can be detected based on image quality characteristics alone. Using both classification and statistical significance testing, the existence and severity of image quality bias in each dataset has been tested and discussed.

Colourfulness

Video footage can undergo massive shifts in perceived colour as both the object composition and lighting can drastically change the perceived colour of a scene. The interpretation of the *colourfulness* of an image is subjective, as an extreme example of this, a colour-blind person has an alternative understanding of the colour of an object due to how their visual receptors process light waves. Spectroscopy, the study of the interaction between electromagnetic waves and matter, can be used to determine the colour composition of a scene based on the distribution of electromagnetic waves emitted from objects that fall within the visible light spectrum. Humans interpret the visible light spectrum in a non-linear manner [25]. Therefore, Hasler *et al.* [55] took a human-centric approach at generating a measure of image colourfulness. The authors performed a qualitative study in which 20 participants were asked to rate the *colourfulness* of a set of images. A metric that correlates highly with human perception was developed to quantify colourfulness. The colourfulness metric defined by Hasler *et al.* [55] was utilised as a feature to distinguish between different action classes in the datasets presented in Section 3.4.

Colour bias arises when one type of action within a dataset has different properties of colour than other footage depicting other actions; this may occur due to hardware or environmental context. The reported Area Under Receiver Operating Characteristic curve AUROC scores obtained using colourfulness as a feature for classification are 0.68, 0.73, and 0.83 for the Violent Flows, UMN, and UCF Web Abnormality datasets respectively. The CF-Violence and NN-Violence datasets achieving an AUROC close to 0.5, suggesting that colourfulness cannot be used to distinguish between *normal* and *abnormal* scenes. In most cases, the Mann-Whitney U test reports that the distribution of colourfulness between classes are significantly different ($p < 0.05$), with the CF-Violence dataset being the only exception ($p > 0.05$). The CF-Violence and NN-Violence dataset contain footage captured by their respective CCTV system during similar times of day/night, resulting in similar colour characteristics and poor classification. The results are in contrast with the UCF and Violent Flows datasets which are composed of footage taken from various sources with different capture characteristics. Normal behaviour is on average more colourful than abnormal behaviour within these datasets.

The Hockey dataset is composed of footage from many different games. However, the environment and general composition of the footage remains fairly consistent, resulting in poor colour based classification. Interestingly, a multi-modal distribution with three unique peaks is observed when analysing the *colourfulness* for the UMN dataset (Figure 3.6). These peaks correspond to the three types of scenes found in the UMN dataset which depict an indoor scene(grey scale), an outdoor scene on grass, and an outdoor scene on concrete. In this case, colour can be used to separate capture devices using the recorded content, but not the action. The UMN case highlights clearly how different environments and hardware can introduce colour bias.



Violent Flows	Hockey	UMN
$U = 1086810$, $p = 0.000$	$U = 4855519$, $p = 0.000$	$U = 78425$, $p = 0.000$
CF-Violence	**NN-Violence**	**UCF**
$U = 355196$, $p = 0.0879$	$U = 7640851$, $p = 0.000$	$U = 64758$, $p = 0.000$

Figure 3.6: Colourfulness distribution for each dataset.

Sharpness

The property of image *sharpness* is a subjective measure that relates to the perceived clarity or detail of an image. Within an image that is considered *sharp*, the boundaries or edges of an object are clear and well defined. Edges within a picture with a low level of sharpness will appear blurred, with the change in tone between two regions within an image being gradual rather than sudden. Figure 3.7 displays the effects of reducing the level of sharpness. The quality of camera lenses and compression techniques can reduce the level of perceived

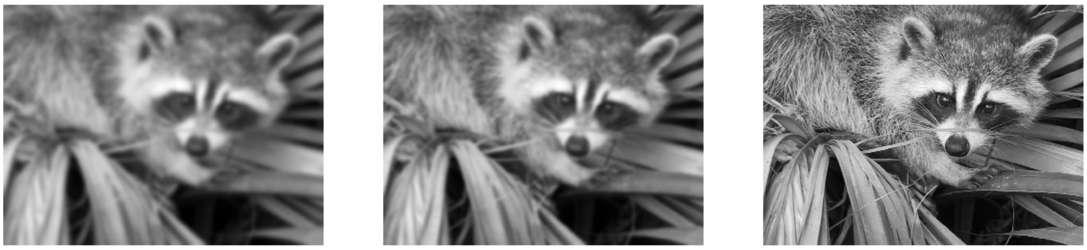


Figure 3.7: The same image with different levels of perceived sharpness.

sharpness.

The perceived sharpness of images extracted from each dataset is measured using the method proposed by Bahrami *et al.* [2]. Bahrami *et al.* introduce a measure of sharpness whose reported values for a set of testing images are closely related to measures of sharpness as reported by humans. Bahrami *et al.* defined the concept of Maximum Local Variation (MLV) as the magnitude of the maximum difference between a pixel and its eight nearest neighbours. The standard deviation of weighted MLV values is used to represent the sharpness of an image.

An AUROC score of 0.91, 0.82 and 0.71 is reported for the CF-Violence, Violent Flows, and UCF datasets respectively (Table 3.1). Regarding these datasets, scenes of normal behaviour depict footage that is, on average, sharper than abnormal (violent) scenes (Figure 3.8). The difference between sharpness between action classes for the NN-Violence, UMN, and Hockey violence dataset are either insignificant (< 0.05) or report a low value; these datasets report a poor classification score. Deniz *et al.* [29] theorised that violent behaviour results in an increased rate of visual blur due to the fast speed of movement associated with a punch and capture characteristics of a camera. This theory may explain the observation that abnormal (violent) scenes in the Violent Flows, CF-Violence and UCF datasets are on average, less sharp.

Deniz *et al.* [29] report a high classification rate when measuring the blur induced by motion when evaluating using the Hockey dataset. Blur is inversely proportional to sharpness, so one would expect measures of sharpness to perform equally well. However, a poor AUROC score is reported. The difference between the two methods is the scope of measurement. The measurement of blur by presented by Deniz *et al.* [29] is locally measured, whereas the measure of sharpness by Bahrami *et al.* [2] is global. Globally, the quality associated sharpness/blur is not sufficient on its own to distinguish between scenes of violence and non-violence when evaluating the Hockey dataset.

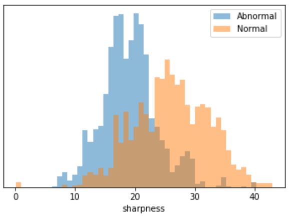
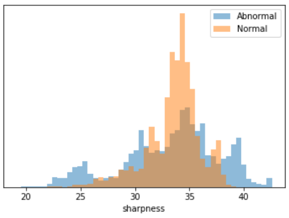
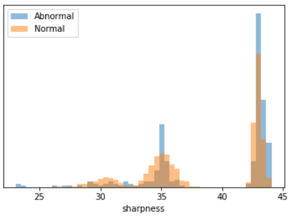
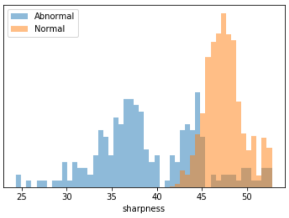
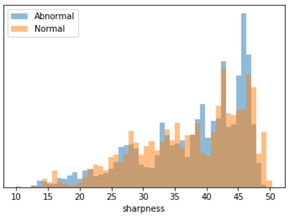
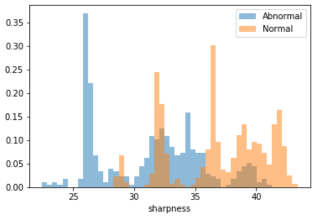
Violent Flows	Hockey	UMN
		
$U = 617348$, $p = 0.000$ $t = -36.58$, $p = 0.000$	$U = 6142590$, $p = 0.000$ $t = 2.87$, $p = 0.004$	$U = 108416$, $p = 0.000$ $t = 3.91$, $p = 0.001$
CF-Violence	NN-Violence	UCF
		
$U = 68198$, $p = 0.0000$ $t = -21.78, p = 0.0000$	$U = 8374255$, $p = 0.1406$ $t = 1.26, p = 0.2060$	$U = 71637, p = 0.0000$ $t = -23.74, p = 0.000$

Figure 3.8: Sharpness distribution for each dataset.

Contrast

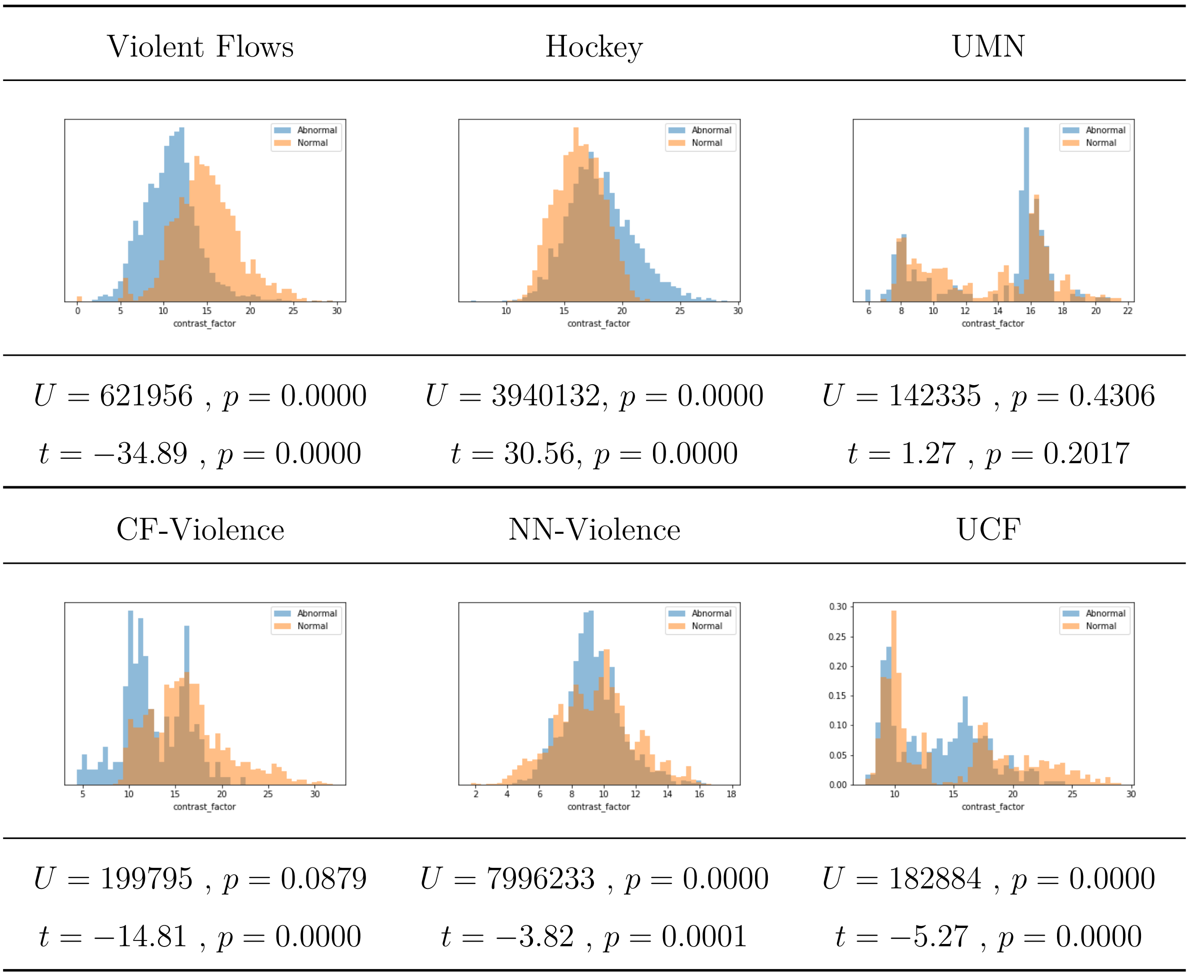
The definition of contrast is the ratio between the brightest and darkest spot in an image. However, according to [96], the human perception of contrast does not entirely comply with this definition. Matkovic *et al.* [96] propose the Global Contrast Factor (GCF) measure of contrast based on the weighted average of local image contrast values obtained at various image resolutions. The local contrast weights are assigned such that the singular contrast measurement align with the human interpretation of contrast of an image. The authors claim that "An image with a high global contrast causes a global feeling of a detailed and variation-rich image. As opposed to it, an image with a lower global contrast contains less information, less details, and appears more uniform".

The contrast characteristic of data that depicts abnormal (violent) and normal scenes is similar for the NN-Violence, UCF, and UMN datasets as reflected by the low t-test and values and AUROC scores that are close to 0.5 (Table 3.1). Except for the UMN and Hockey datasets, the contrast factor for abnormal scenes is lower on average than data depicting normality (Figure 3.9).

Complexity

Image complexity refers to the perceived intricacy of patterns and is believed to have a substantial impact on aesthetic appreciation [89]. Machado and Cardoso present a method for computing an aesthetically directed value of complexity [88]. An estimate of image complexity is calculated by measuring the Root Mean Square Error (RMSE) between an image before and after applying JPEG compression; the compression ratio then divides this value. The concepts of compression and complexity are related [116], and that images with high complexity are poorly represented when subject to high, lossy compression.

Complexity does not necessarily capture characteristics associated with the cap-



Violent Flows	Hockey	UMN
$U = 621956$, $p = 0.0000$ $t = -34.89$, $p = 0.0000$	$U = 3940132$, $p = 0.0000$ $t = 30.56$, $p = 0.0000$	$U = 142335$, $p = 0.4306$ $t = 1.27$, $p = 0.2017$
CF-Violence	**NN-Violence**	**UCF**
$U = 199795$, $p = 0.0879$ $t = -14.81$, $p = 0.0000$	$U = 7996233$, $p = 0.0000$ $t = -3.82$, $p = 0.0001$	$U = 182884$, $p = 0.0000$ $t = -5.27$, $p = 0.0000$

Figure 3.9: Contrast Factor distribution for each dataset.

ture quality of a camera system but rather the content structure. Therefore, a reasonable assumption would be that footage obtained from a single image capturing system would have similar global characteristics. Using city centre based CCTV as an example, crowd compositions would be expected to be visually comparable regardless of local actions and behaviours. If this is false, then there exists the possibility that data has been gathered in a way such that bias is introduced. In the case of crowds, *complexity bias* manifests when footage of violence is captured in a different manner than footage of non-violence, this can occur due to camera operator movement and zoom. With this in mind, evaluating the CF-Violence dataset using image complexity reveals perfect separation between violent and non-violent scenes (Table 3.1). Ideal classification using complexity suggests that the global composition of pedestrians in violent scenes are entirely different to pedestrians depicted in scenes of normality. An AUROC of 0.85 (Table 3.1) is reported when evaluating the Violent Flows dataset, suggesting that the visual composition of violent and non-violent crowds are different. The remaining datasets reported weak AUROC scores, indicating that abnormal and normal scenes are visually similar. With the exception of the UMN dataset, images that depict abnormal behaviour are typically considered less visually complex than normal behaviour as indicated by statistically significant (< 0.05) t-test t values (Figure 3.10).

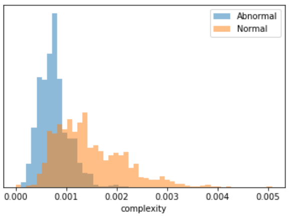
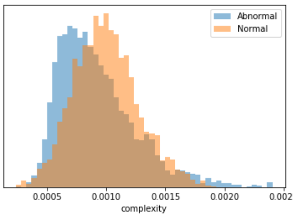
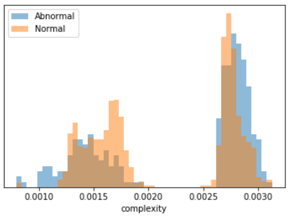
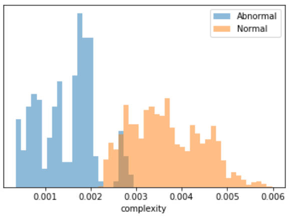
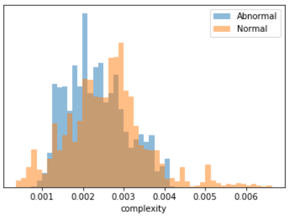
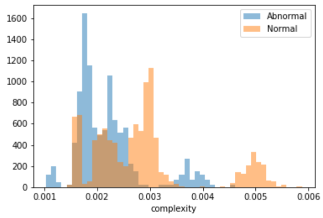
Violent Flows	Hockey	UMN
		
$U = 506425$, $p = 0.0000$ $t = -38.84$, $p = 0.0000$	$U = 5009512$, $p = 0.0000$ $t = -10.31$, $p = 0.0000$	$U = 122321$, $p = 0.0000$ $t = 3.23$, $p = 0.0001$
CF-Violence	**NN-Violence**	**UCF**
		
$U = 7080$, $p = 0.0000$ $t = -50.25$, $p = 0.0000$	$U = 7109462$, $p = 0.0000$ $t = -12.74$, $p = 0.0000$	$U = 7640851$, $p = 0.0000$ $t = -5.27$, $p = 0.0000$

Figure 3.10: Complexity distribution for each dataset.

Estimated Noise

Image noise manifests itself as random variation in pixel intensity caused by electronic noise, usually affecting the imaging sensor [80]. The quality of image capture hardware is typically related to the amount of captured noise, and poor quality hardware is likely to produce noisy imagery. The variance of additive Gaussian noise was estimated using a method proposed by Immerkaer [66]. Their approach is reportedly insensitive to the image structure. The author highlights that thin lines may be considered to be noise by their estimation algorithm "In highly textured images or regions, though, the noise estimator perceives thin lines as noise.".

The distribution of noise for each action class are statistically different in both distribution shape and mean (< 0.05). However, the t-test value for the UMN and UCF datasets are low. On average, scenes of abnormal behaviour contain less noise when compared to images of normal behaviour. The noise characteristics allow for near perfect classification for the CF-Violence dataset (AUROC $= 0.99$), indicating that violent and non-violent scenes have different image quality characteristics. Considering that the noise estimation may interpret thin lines as noise, the high classification scores for the CF-Violence and Violent Flows datasets may arise from dense crowds being considered noise.

Full-Reference Quality Analysis

In this section, full-reference image quality metrics are described, and their ability to generate a model that can distinguish between violent and non-violent behaviour is evaluated. The methods discussed in this section operate by applying a transformation to an image and measuring the difference between the original and transformed image. Each approach to full-reference quality analysis is defined by the particular way in which the difference between the original and transformed

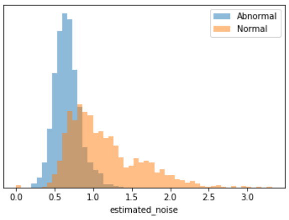
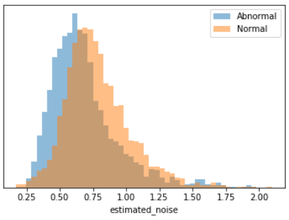
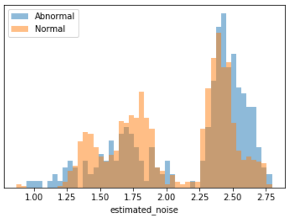
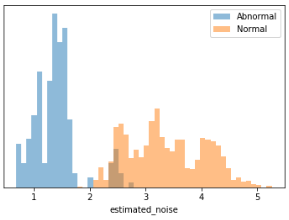
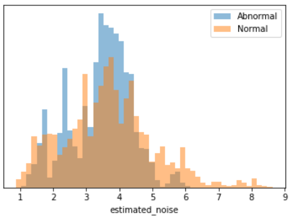
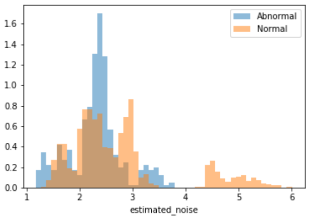
Violent Flows	Hockey	UMN
		
$U = 453528$, $p = 0.0000$	$U = 4480915$, $p = 0.0000$	$U = 112382$, $p = 0.0000$
$t = -39.39$, $p = 0.0000$	$t = -16.51$, $p = 0.0000$	$t = 4.39$, $p = 0.0001$
CF-Violence	NN-Violence	UCF
		
$U = 5347$, $p = 0.0000$	$U = 7874664$, $p = 0.0000$	$U = 164904$, $p = 0.0000$
$t = -67.63$, $p = 0.0000$	$t = -8.82$, $p = 0.0000$	$t = -10.51$, $p = 0.0000$

Figure 3.11: Estimated Noise distribution for each dataset.

image is measured. Three methods of full-reference IQA are evaluated; these are Peak Signal-to-Noise Ratio (PSNR) [148], Visual Information Fidelity (VIF) [120] and *compression ratio.*

The transformation applied to an image extracted from a video is image compression. Specifically, Joint Photographic Experts Group (JPEG) compression is used to transform an image for comparison with the uncompressed original. JPEG compression within the Scikit-image Python library [138] accepts a quality parameter whose potential input is an integer within the range $[0, 100]$. The quality parameter increases the rate of compression, higher values sacrifice image quality in exchange for a smaller file size.

Using these measures, the full-reference quality of a single frame within a video is represented using a feature vector where each element represents a full-reference measure computed using a different image transformation; in this case, the difference in transformation is controlled by changing the rate of JPEG compression. Compression is performed multiple times using 50 compression values starting at 0 and increasing at an interval of 2. Computing a measure for all 100 compression rates was computationally expensive. A single frame in a video is therefore represented by a feature vector of 50 values where each element is a full-reference quality measure computed using compression at a given level.

Full-reference methods and results

Peak Signal-to-Noise Ratio (PSNR) (Equation 3.1) is used as a quality measurement between an image and its compressed counterpart, greater values indicate that the applied compression preserved detail and maintains quality. The output of PSNR does not correlate with the human perception of image quality [148].

$$\text{PSNR} = 10\text{log}_{10}(\frac{R^2}{\text{MSE}}) \tag{3.1}$$

Sheikh *et al.* [120] proposed the Visual Information Fidelity method of full-reference IQA. This method is targeted towards quantifying the quality of *naturalistic scenes* and performing human-like quality assessments. The Visual Fidelity measure is a ratio between two full-referential quality measures, the amount of reference information that can be extracted from a distorted image as informed by natural scene statistics, human visual system, and a distortion model. The second measure of quality used in the ratio is the amount of shared information between the reference and distorted image [121].

Each of these methods reports similar results when performing quality based classification for abnormal scene detection. In each case, the reported classification metric suggests greater than random selection when determining whether a scene is normal or abnormal (violent). The CF-Violence dataset achieves almost perfect classification performance and the Violent Flows and Hockey Violence dataset report high classification. The UMN and NN-Violence datasets are classified least effectively using full-reference IQA.

Table B.1 (Appendix B) presents the statistically significant (< 0.05) correlation values between each full-reference binary predictions and no-reference IQA metrics. In general, the full-reference IQA does not correlate strongly with colourfulness, suggesting that the transformation induced by JPEG compression has little effect on the full-reference IQA methods. In the case of CF-Violence, strong correlations with estimated noise, contrast, sharpness, complexity and contrast factor are observed, suggesting that the full-reference IQA approach may be describing, in some part, these aspects of quality. Regarding the instances of low, or insignificant correlation between no-reference metrics and full-reference IQA, It can be inferred that the full-reference IQA are potentially focusing on other factors of quality not explicitly investigated.

Image Quality Assessment Discussion

In this section, we will discuss the potential effects of IQA based biases within each dataset. Averaging the classification score obtained using each IQA method provides a general indication of the image quality based bias for each dataset. The datasets ordered in descending average AUROC scores are: CF-Violence, Violent Flows, UCF, Hockey, UMN and the NN-Violence, each reporting average classification performance of 0.89, 0.83, 0.73, 0.69, 0.62 and 0.60 respectively.

The appearance of footage taken from the Violent Flows and CF-Violence datasets have distinct characteristic differences between violent and non-violent samples. These different characteristics may arise through the physical act of capturing the data. Camera operators target scenes of interest and attempt to focus on the ac-

Feature Type	Violent Flows	Hockey	UMN	CF-Violence	NN-Violence	UCF
estimated_noise	0.87 ± 0̇37	0.64 ± 0.019	0.6 ± 0.139	0.99 ± 0.015	0.53 ± 0.052	0.63 ± 0.253
sharpness	0.82 ± 0.052	0.49 ± 0.006	0.62 ± 0.153	0.91 ± 0.098	0.49 ± 0.011	0.78 ± 0.278
colourfulness	0.68 ± 0.045	0.55 ± 0.105	0.73 ± 0.082	0.51 ± 0.092	0.55 ± 0.088	0.83 ± 0.185
complexity	0.85 ± 0.034	0.6 ± 0.02	0.55 ± 0.156	1.0 ± 0.007	0.58 ± 0.029	0.73 ± 0.174
contrast_factor	0.81 ± 0.05	0.69 ± 0.011	0.45 ± 0.123	0.74 ± 0.181	0.53 ± 0.049	0.5 ± 0.232
vifp	0.89 ± 0.016	0.88 ± 0.009	0.55 ± 0.087	1.0 ± 0.003	0.76 ± 0.067	0.74 ± 0.197
psnr	0.9 ± 0.019	0.88 ± 0.02	0.79 ± 0.069	1.0 ± 0.001	0.73 ± 0.074	0.72 ± 0.263
size	0.8 ± 0.046	0.79 ± 0.012	0.66 ± 0.076	0.99 ± 0.013	0.65 ± 0.052	0.9 ± 0.059
average	0.83	0.69	0.62	0.89	0.60	0.73

Table 3.1: Classification performance (AUROC) demonstrating the ability to distinguish between violent and non-violent scenes using image quality for the Violent Flows dataset.

tion by changing the camera angle and zooming in of the region of interest. In the case of Violent Flows and CF-Violence, the perceived sharpness of violent scenes are lower on average than scenes of non-violence. One hypothesis is that camera operation may influence this. The act of changing the camera angle to best capture a scene may induce a motion blur. Applying digital and analogue camera zoom can alter the quality characteristics of captured footage. Digital zoom will reduce the perceived resolution whereas analogue zoom can induce visual distortions. In addition to this, when zoomed in, the camera operators movements will be accentuated, increasing the probability of causing a motion blur. Reduced image sharpness is associated with reduced contrast, noise and image complexity (Appendix C Tables C.1 – C.6). An alternative hypothesis is that the structures observed in normal scenes are different to those in abnormal scenes. It may not be the operator motion that induces changes in quality characteristics, but rather that the difference in data capture represents similar structures, such as pedestrians, from different perspectives. This assumes that the quality measurement approaches describe the structures of an image rather than aspects of quality. The method of noise estimation adopted is reportedly insensitive to the structure of the image, however estimated noise correlates strongly (Pearson's R > 0.9) with complexity which is not insensitive to image structure (Appendix C Tables C.1 – C.6). This may suggest that either noise estimation identifies the structure as noise and complexity identify noise as image structure or vice versa; it is also possible that both are true.

The IQA measures used are globally descriptive; therefore if a scene is similar in composition to normality with only a local difference of a person acting locally violent or abnormal, then one would expect the distribution of image quality statistics to be similar for normal and abnormal scenes. This was observed when analysing the sharpness of the Hockey dataset. Our global evaluation of sharpness yields poor classification ability whereas the local approach by Deniz *et al.* [29] demonstrates greater classification capacity. The fact that the CF-Violence and

Violent Flow datasets achieve good classification ability using IQA may suggest that the appearance of the two classes of action are highly dissimilar in either quality or scene composition.

Creating invariant systems is preferable as they, in theory, generalise well. Identifying methods that normalise quality characteristics will aid machine learning in the avoiding solving tasks using data that can be considered *incorrect* or *irrelevant* to the problem. Although a weak indicator of abnormal behaviour, the image quality of colourfulness is easy to remove. The potential impact of colour can be removed by converting images to greyscale. Other factors such as contrast and sharpness are less easily normalised. The sharpness of an image can be adjusted using image filters such as a sharpening or blur filter. Data can be theoretically tuned to have similar sharpness characteristics. However, this process may remove valuable data or introduce incorrect information after applying a blur or sharpen filter respectively. Similarly, the contrast of an image can be balanced using histogram normalisation procedures [164]. As with sharpening, contrast normalisation may affect other qualities such as perceived noise. Data augmentation is standard practice within the Deep Learning field of computer vision. Augmenting data to display various quality characteristics may prove beneficial when generating an invariant model.

3.5.2 Understanding Depth

Intuitively, depth invariance could be understood to be universally preferable; a highly invariant system is robust to biases in the data and has more potential as a general solution to a subset of problems. Depth invariance is very important when considering surveillance systems. It is often better to deploy a camera system that maximises the amount of data captured. Maximising data capture is often realised by placing cameras at high altitudes, with the focal direction pointing down the length of a street. Using this setup, objects within the street exist at various depth levels from the camera, and due to the projection of a 3D scene to a 2D image, a particular object placed at a variety of positions along the street has different sizes or scales, relative to its depth.

Using work by Laina *et al.* [76], the approximate perceived depth of a scene can be obtained. The work mentioned above attempts to approximate a depth map from a single image using a deep residual network pre-trained on the NYU Depth dataset [125]. The depth of a frame is encoded using five statistics; mean, median, min, max and standard deviation of the predicted depth map. 5-fold cross validation and a linear SVM classifier are used to obtain classification scores that dictate the ability to separate samples of violence and non-violence based on the depth derived statistics.

The Hockey and CF-Violence demonstrate the mode of data capture has introduced a depth based bias into each dataset, see Table 3.2. Hockey is a team based spectator sport, and television broadcasts benefit from a camera operator focusing tightly on an event of interest. In a standard hockey game, to observe a pass of the puck from one player to another, a wide camera shot would be used as distances between connecting players can be large. In contrast, when a fight occurs, the proximity between players is small, and so a close-up, tight fitting camera shot better portrays the actions of interest to the spectators. On a similar note, the violent data contained in the CF-violence dataset is centred and focused. The camera operator had identified that a fight was underway and tightly fit the camera to the action to capture identifiable features of the perpetrators. This explanation holds true for the NN-violence dataset, however to a lesser degree. Applying this process to the Violent Flows dataset shows classification performance metrics that represent a minor depth bias.

Dataset	ROC AUC	Accuracy ±
CF-Violence	0.92	84.2 ± 2.6
NN-Violence	0.70	63.5 ± 2.1
Hockey	0.73	68.6 ± 4.1
Violent Flows	0.62	55.3 ± 2.0

Table 3.2: Classification when identifying violence using measures of depth.

3.6 Conclusion

Perceived image quality and depth are two factors that can be exploited to achieve state-of-the-art classification score on violent behaviour datasets. Depth and quality bias in the data is problematic as facets do not convey action, a key determinant of violence. An unsupervised learning technique may generate a solution that learns image quality/depth statistics rather than motion analysis based features. The study in this section has highlighted the potential problems by experimentally demonstrating that good classification performance is achieved using information that could be considered irrelevant to the task.

To avoid classifying a scene by its perceived depth or image quality characteristics, and instead judge a scene based on the contained actions, it is beneficial to develop methods that have inbuilt depth/quality normalisation techniques. Unfortunately, normalisation may not be enough. Due to the data collection process, existing datasets do not represent action classes equivalent in all aspects but the action (violent/non-violent). In the case of CCTV, this can bring into question the reported ability of a method to correctly classify data. As previously stated, in the case of the CF-violence dataset, violent scenes are typically tightly focused on the event whereas non-violent scenes are usually wide, long shots. By tightly focusing on violence, information associated with a street environment such as pedestrians walking and interacting are removed. The lack of surrounding context presents a task of “detect violence”, and not the task of, “given a street environment, detect violence.” Although similar, these tasks are not equivalent. The latter task requires a solution that can distinguish between regions pertinent to violence from other regions associated with non-violent behaviour.

Chapter 4

Modelling Crowds using Temporal Texture

4.1 Introduction

Datasets presented in Chapter 3 are characterised as binary classification datasets, the ground truth defines each video as *violent or non-violent*, or *normal or abnormal.* The available datasets lack both transitional information that depicts the actions that occur between normality and abnormality. Therefore, violent behaviour prediction is not attempted due to the lack pre-violence information with which to formulate a prediction model. Additionally, the severity of violence within each dataset has not been evaluated, removing the potential to apply a regression analysis technique to evaluate the severity of an incident. Due to these limitations, violent crime detection as presented in this thesis is treated as a binary classification problem; Given existing examples of violence, fit a model to evaluate an unseen sample to determine its nature, violent or non-violent. To perform this process, a feature representation that encodes important aspects of data must be obtained. This representation should describe characteristics of violence and non-violence while avoiding other aspects such as image quality (See Chapter 3). Violent behaviour can manifest in many different forms, with two distinct classes prominent in the data, these being crowd violence, and one-on-one violence. Within city centre environments, the pedestrian population is high, resulting in the emergence of crowding. Measures of image texture are well suited for describing the seemingly unstructured patterns that result from the mass occlusions caused by crowding [91]. Violent behaviour is defined partly by the actions and movements of engaged combatants, therefore modelling the temporal dynamics can be useful. Unfortunately, traditional optical flow approximation methods perform poorly on crowded scenes due to self-occlusion [72, 92]. The work presented in this chapter is based on the assertion that the appearance of abnormally behaving crowds will undergo different patterns of change when compared to crowds exhibiting normal behaviour. Therefore, a description based on encoding the change in crowd appearance over time is proposed.

Introduced in this Chapter is the Inter-Frame Uniformity (IFU) measure, a value that quantifies the uniformity of a sequence that when applied to a visual descriptor, describes the stability of a crowd's appearance over time. This measure was proposed by looking at the data. Generally, the appearance of violent crowds between successive frames depicts greater changes in appearance when compared to normal behaviour whose appearance changes more gradually. In addition to this, it was observed that the rate of change of appearance was more variable for violent behaviour; the appearance of normal behaviour changes in a more consistent manner. Based on observation, a hypothesis was formulated and tested, that *over a short-time period, the appearance of violent behaviour is less stable over time when compared to footage of non-violent behaviour.* During the testing of this hypothesis, it was demonstrated that IFU is a powerful descriptor for use in a violent behaviour classification pipeline.

It has been argued that many methods of violent behaviour detection are too computationally expensive to be practically implemented in the real-world [49]. As discussed in Chapter 1 & Chapter 2, Section 2.1, theoretical benefits to public health and safety can be achieved by using computer vision to assist in the *active observation* of surveillance footage. With this in mind, the algorithm presented in this chapter was designed to operate with a low computational cost.

To summarise, a computationally cheap method of abnormal crowd description that achieves state-of-the-art results across many datasets, including real-world CCTV dataset known as CF-Violence, is presented. The proposed method generates a scene description that can be used to discriminate between abnormal and normal scenes in the *UMN unusual crowd*, and *UCF Web abnormality* datasets. State-of-the-art discrimination between violent and non-violent scenes as shown in the *Violent Flows* and *CF-Violence* datasets is also reported. An extensive investigation of the parameter effects of the proposed method is presented. It is demonstrated that violent behaviour has the property of non-uniform change over time.

4.2 Related Work

Violent behaviour detection in crowded situations can be considered a subset of the *abnormal crowd detection* field of research. For this field, a vast selection of approaches exists in the literature. Within the literature, there exists an assertion that using optical-flow approximation to uncover motions within dense and complex crowds is infeasible as flow approximations are poor [72, 92]; this has affected the design of feature representations used to identify abnormally behaving crowds. Kratz *et al.* [72] avoid optical flow based motion description by extracting fixed size spatiotemporal volumes and computing spatiotemporal gradients of pixel intensities which are represented using a three-dimensional Gaussian Mixture Model (GMM). The authors model normal behaviour using a Hidden Markov Model and declare a new observation as abnormal if it does not fit the learnt model. Wang *et al.* [145] also avoid an optical flow based representation when dealing with crowds, they instead favour statistics computed from wavelet transformed spatiotemporal slices taken from a spatiotemporal volume. Although the effectiveness of optical flow is often a point of theoretical contention when dealing with crowds, there exist many methods that utilise measures of optical flow with excellent results. Ryan *et al.* [115] encode optical flow vectors using a three-dimensional GLCM structure, expressing the dynamics of a local region by the texture of motion. The authors generate a model of normality using a GMM. The authors claim that their method is both effective at discerning between normal and abnormal scenes while maintaining an arguably real-time processing speed of approximately 9 FPS. Their approach would require input data to be temporally down-sampled to 9 FPS for real-time use, sacrificing potentially important information for speed. Wang *et al.* [146] compute global Histogram of Optical Flow Orientation (HOFO) on a per-frame basis and model normal behaviour using two separate methods, one-class SVM learning, and Kernel-Principal Component Analysis (PCA) embedding. The authors show that the two approaches are effective at modelling normal behaviour when used in conjunction with their proposed descriptor; however, the one-class SVM offered slightly higher performance. Chen *et al.* [17] extracted a notion of crowd acceleration and stated that rapid changes in acceleration could be used to identify a crowd displaying normal activity from a crowd currently undergoing a situation of panic. Recent work by Biswas *et al.* [6] express the problem of abnormal behaviour analysis as identifying sparse, or rarely occurring behaviours. A matrix of features represent video frames, and matrix decomposition is applied to separate the matrix components into two groups, low-rank and sparse components; the latter of which is considered anomalous.

An alternative approach for modelling motion involves tracking features to obtain motion trajectories. Zhou *et al.* [160] train a Multi-Observation Hidden Markov Model (MOHMM) using motion trajectories extracted from footage of normal behaviour using a KLT feature tracker. The probability of the trained model producing a given observation is computed, if the probability falls beyond a threshold, then the observation is considered abnormal. Marsden *et al.* [93]

utilise a KLT tracker to extract motion trajectories and compute holistic measures of crowd collectiveness, conflict, and density. These measures form a feature vector that describes the dynamics of a scene. The approach described by Zhou *et al.* [160] applies to many domains of crowd abnormality as it does not assume any specific measurement of crowd motion, whereas the holistic approach by Marsden *et al.* [93] is useful on data where the measures documented are known to exist. Although tracking has shown to perform well at describing crowd behaviour, Yang *et al.* [155] highlight the difficulty in tracking when analysing scenes with changing illumination, a property common to naturalistic environments.

Early research by Marana [91] formulated the crowd density estimation problem as a global measure of visual texture. Marana showed that sparse and dense crowds hold notably different textural compositions. Multiple studies [3, 16, 21, 91, 142] utilised the GLCM approach for crowd description and had shown that Haralick's GLCM features could be used to determine the density of a crowd successfully; the implication being that texture can provide a meaningful description of the visual appearance of crowds. Rao *et al.* [112] identified the usefulness of GLCM for surveillance analysis and incorporated a GLCM description of tracked pedestrians as part of their anomalous behaviour detection framework. To justify the choice of GLCM as a spatial descriptor, the authors refer to a crowd counting paper that utilised GLCM [16]. However, this method is reliant upon tracking objects, an operation that performs poorly on highly populated crowds or scenes with large amount of occlusion, like those found in city centre CCTV. In contrast with the work presented by Rao *et al.* [112], the work presented in this chapter focuses on describing how the appearance of a crowd changes over time as opposed to how local individuals change over time. Global descriptors report state-of-the-art performance for the task of violent behaviour detection in crowds while operating in real-time [56, 85]. Specific descriptor methods are capable of achieving similar performance on crowd datasets [49, 104], however, this usually incurs a high computational cost resulting in systems that typically fails to operate in real-time [49]. Presented in Chapter 5 is a real-time local descriptor targeted towards the analysis of one-on-one violence [84]. To analyse crowds in real-time using the local descriptor approach, GPU hardware acceleration is required. In contrast to this, the work presented in this chapter operates in real-time on standard hardware.

4.3 Proposed Method Overview

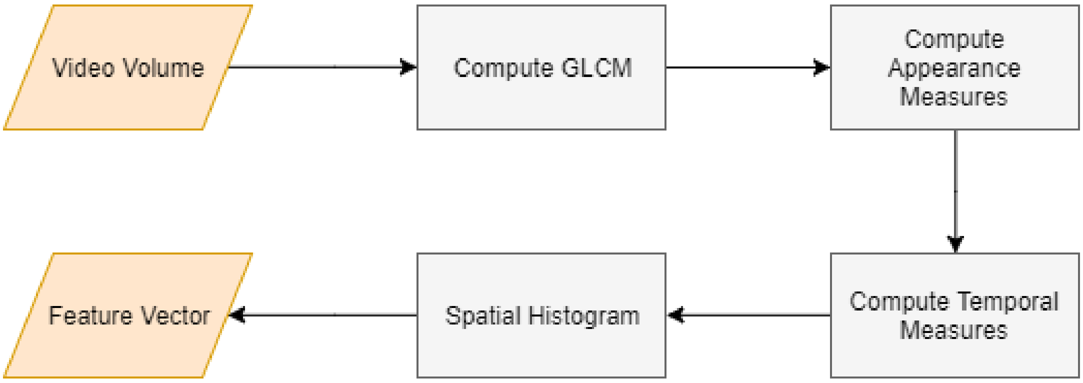


Figure 4.1: Flow chart for the proposed method.

The proposed method builds upon Haralick texture features [53] which describe visual texture using statistics derived from co-occurring grey level intensities. Haralick features are computed for each frame in a sequence. By describing how Haralick features evolve over time using simple summary measures, a succinct and powerful descriptor of crowd dynamics that yields fast compute time and robustness to change over time, is generated. Haralick texture features are extracted from a GLCM. Haralick texture features are extracted from a GLCM. A GLCM is generated by counting the co-occurring grey level intensity values found in an image given a linear spatial relationship between two pixels. The spatial relationship is defined by a parameter pair (θ, d) where θ is the orientation and d is the distance between two pixels. It is common to define a set of parameter pairs (θ, d) and to then combine GLCM matrices, this is typically used to provide rotational invariance by using a set of orientation parameters, typically in 8 orientations, spaced $\pi/4$ radians apart. The number of grey level values N_g represents the number of unique intensity values present in an image. It is common to scale an image from $[0, 255]$ to $[0, N_g]$ before computing a GLCM, where N_g is a defined number of gray-levels [53].

The following features as defined by Haralick [53] are computed: *Energy*, *Contrast*, *Homogeneity*, *Correlation* and *Dissimilarity*. These measures were identified to be statistically important [54] and as a result, are implemented in many code libraries and toolkits. The variable $P(i, j)$ expressed in Equations (4.1-4.5) refers to the value at the $(i, j)^{th}$ position in a gray level co-occurrence matrix.

$$\text{Angular Second Moment} = \sum_{i,j=0}^{N_g-1} P_{i,j}^2 \text{‘} \tag{4.1}$$

$$\text{Contrast} = \frac{\sum_{i,j=0}^{N_g-1} P_{i,j}(i-j)^2}{(N_g-1)^2} \tag{4.2}$$

$$\text{Homogeneity} = \sum_{i,j=0}^{N_g-1} \frac{P_{i,j}}{1+(i-j)^2} \tag{4.3}$$

$$\text{Correlation} = \frac{\left[\sum_{i,j=0}^{N_g-1} P_{i,j}\left[\frac{(i-\mu_i)\ (j-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}}\right]\right]+1}{2} \tag{4.4}$$

$$\text{Dissimilarity} = \frac{\sum_{i,j=0}^{N_g-1} P_{i,j}|i-j|}{N_g-1} \tag{4.5}$$

The equations for contrast (Equation 4.2), correlation (Equation 4.4) and dissimilarity (Equation 4.5) are scaled such that the returned value is bounded between $[0, 1]$. Given a series of images expressing appearance over time, the aforementioned texture features are computed in order to generate a time ordered sequence of texture over time, referred to as x. A statistical summary of each sequence is calculated to encode the underlying crowd behaviour. Each sequence x is represented as a four length feature vector composed of measures of mean, standard deviation, skewness (Equation 4.6) and inter-frame uniformity (IFU, Equation 4.7). Skewness indicates the asymmetry found in a distribution and can be used to deduce whether a distribution is showing a general increase or decrease in value over time. Inter-frame Uniformity (IFU) as expressed in Equation 4.7 is a measure of adjacent sample similarity within time ordered data. It is expressed as the scaled L_2 norm (Equation 4.7) of the sequence y where sequence y is formed by taking the absolute difference between adjacent samples within sequence x, $y_t = |x_t - x_{t+1}|$. Sequence y is normalized by its sum before being input into Equation 4.7. IFU returns values within the range $[0, 1]$ where 0 and 1 represent non-uniform, and uniform change over time respectively. This particular measure of uniformity was designed to be sensitive to sudden change in value over time and is therefore intended to be suited towards highlighting more abrupt changes in time-ordered data. As discussed in the introduction of this chapter, IFU was created after observing the available data. It was observed that scenes of violent behaviours in crowds are characterised by a substantial and inconsistent rate of change in appearance over short periods. IFU provides a measure with which to test whether the observed property truly characterises violent crowds.

$$\text{Skewness} = \frac{E(x-\mu)^3}{\sigma^3} \tag{4.6}$$

$$\text{IFU} = \frac{|y|_2\sqrt{(T-1)} - 1}{\sqrt{(T-1)} - 1} \tag{4.7}$$

It was observed that different spatial regions in frame depicted different behaviour, therefore each video is spatially sub-divided into $M \times N$ non-overlapping sub-regions (cells) before applying the aforementioned method to each. Each cell is represented by twenty values that describe the appearance, and change in appearance over time. Twenty histograms are generated using values taken from each cell within the $M \times N$ grid, one histogram for each feature. Through empirical analysis it was found that using logarithmically distributed histogram bins within the range of $[0, 1]$ provided the best performance. Histograms representing skewness are bounded between $[-1.4, 1.4]$ and the bins are logarithmically, and symmetrically distributed around zero such that bin spaces are closer at values closer to zero.

In the case of surveillance footage, failure to remove background information may lead to the description of landmarks as opposed to crowd dynamics. Two GLCMs generated adjacent in time will have a very similar composition as static objects will introduce the same information in both matrices. To remove static information that typically corresponds to the background of a scene, adjacent GLCMs are subtracted $M_t - M_{(t-1)}$, where t represents the frame being analysed. All values less than zero are assigned a value of zero. This approach comes at a near negligible computational cost and offers robustness to minor translational camera motion due to the spatially unconstrained nature co-occurrence matrices.

4.4 Experiments and Results

As discussed in the introduction to this chapter, the violent behaviour detection problem is formulated as a binary classification task. For each frame of a video, an algorithm will report whether violent behaviour or non-violent behaviour is depicted. In this section, an overview of the classification methodology, experimental set-up and results are presented.

A classification label is generated for each video frame in order to provide a continuous activity feed usable in CCTV observation scenarios. This is achieved by classifying a description vector computed using the previous n frames in sequence, by default n is assigned to be equal to the number of frames per second. The parameter N_g, used to generate the grey level co-occurrence matrix, is assigned a value of 32. The parameters (θ, d) are assigned as $(0, 1)$, see Section 4.4.1 for an explanation regarding the choice of these parameters. M and N, which specify frame sub-division used to encode spatial information are assigned the value of 4, Section 4.4.2 discusses the effects of using different values for M and N. All experiments were conducted using K-fold cross validation where data is split into K partitions with $K - 1$ partitions being used for training a random forest classifier [11]. The remaining partition is used for testing; the random forest is composed of 50 trees. The parameter K is assigned a value of 5, 5, 2, and 2 when processing the CF-Violence, Violent Flows, UMN and Web Abnormality datasets respectively. The choice of $K = 2$ was informed by the literature [98]. To reduce variability introduced by random sampling during cross validation, each experiment is performed 10 times and the average result is reported. As stated previously, features are extracted such that each frame in a sequence is represented by a single vector. Features extracted from a single source video are not permitted to be placed in both training and testing partitions at the same time, as features extracted from any single video are likely to belong to the same distribution and may lead to over-fitting.Results are reported using receiver operating characteristic (ROC) curves, A common way to summarise these curves is to report area under the curve. Area under ROC dictates the discrimination performance between binary classes, a value of 1 indicates perfect discrimination. The proposed method was implemented using Python and the Scikit-image library. All experiments were performed using an Intel i7-4790 at 3.6GHz processor. Given a temporal window size n of 24, and a resolution of 640×480, the proposed method operates at 76.92 frames per second, or 0.013 seconds per frame.

Decomposing the importance of temporal features it was found that the measure of Intra-frame Uniformity is highly descriptive (Figure 4.2) when applied to the two datasets whose *abnormal* class contains only violent samples, these being the CF-Violence and Violent Flows datasets. Looking at the average IFU values returned by these datasets, it was observed that the appearance of violent scenes within the Violent Flows dataset change in a less uniform manner over time (Table 4.1). It was also observed that appearance of scenes in the CF-Violence dataset, as represented by ASM and Homogeneity measures, exhibit the same property. Given this observation, an additional experiment was formulated to deduce whether or not a lower IFU is indicative of violent behaviour when compared to normal behaviour. The Web Abnormality dataset contains examples of violence within the *Abnormal* class. All non-violent abnormal scenes have been separated to create two new binary datasets, these are *Violent* or *Normal* (VoN), and *Violent* or *Abnormal* (VoA), the latter differs in that the *Abnormal* class is composed of the Non-Violent abnormal samples from the Web Abnormality dataset. It was observed that the IFU measure reported across all appearance features for both VoN and VoA (Table 4.1), is less for scenes of violence, this suggests that violence has a greater non-linear change in appearance over time when compared to non-violence. Continuing the IFU analysis, it was found that scenes of abnormality as displayed in the UMN dataset have a greater IFU than scenes of normality, this highlights that a low IFU is not indicative of all types of abnormal behaviour.

When testing the UMN dataset, the proposed approach achieves comparable classification ability to other state-of-the-art methods (Table 4.3) when using a temporal window size greater than 64 frames in length (Table 4.7). The results show that the rate of classification increases as the temporal window length increases. It is believed that as the panic situation winds down, the key characteristics of panic are less prominent, therefore increasing the temporal window size prolongs the time in which the dominant characteristics remain in the decision making process.

The results reveal that the measure of mean appearance over time is the a weakest descriptor when applied to the Web Abnormality and Violent Flows datasets (Figure 4.2), it is hypothesise that the appearance of crowds within these data-

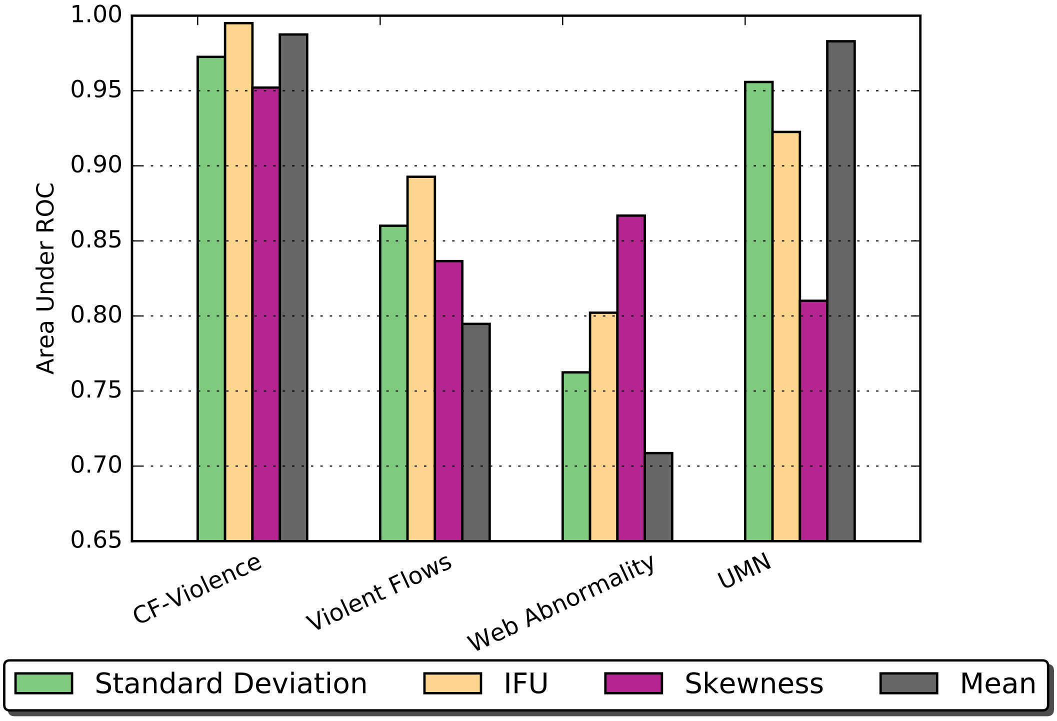


Figure 4.2: Classification performance achieved by each temporal feature type.

sets vary a lot as the are recorded from different sources, therefore strong visual correspondences in appearance across samples is unlikely. Both the UMN and CF-Violence dataset use fixed cameras which record different crowd behaviours within the same environment. Given that the environment can guide the flow/behaviour of a crowd, then typical crowd compositions emerge during scenes of normality, in which case the mean appearance offers high classification ability as inter-sample visual similarity is more likely to occur between samples that depict normality. An evaluation on the Violent Flows dataset using the methodology outlined in the seminal paper by Hassner *et al.* [56]. The proposed method offers comparable performance with existing methods (Table 4.5). The only alternative method to report a greater classification accuracy than the proposed approach does not operate in real-time [49], unlike the proposed method which does boast real-time performance.

Table 4.1: Inter-Frame Uniformity measure difference between scenes of Abnormality and Normality for each texture measure outlined in Section 4.3. Negative values indicate Normal scenes have a greater value than Abnormal scenes, indicating that scenes of Normal behaviour are temporally more uniform. (T-test, * <0.05, ** < 0.01).

IFU	Dissimilarity	Correlation	Homogeneity	ASM	Contrast
CF-Violence	0.0564**	-0.0731**	-0.0325**	-0.0316**	0.0773**
Web Abnormality	0.0382**	-0.0183**	-0.0435**	-0.0334**	0.0246**
UMN	0.0318*	0.0290	-0.0421**	0.0539*	0.0500**
Violent Flows	0.0257**	-0.0287**	-0.0206**	-0.0690**	0.0232**
VoA	0.0247**	-0.0049	-0.0105**	-0.0319**	0.0128**
VoN	0.0331**	-0.0098	-0.0221**	-0.0381**	0.0186**

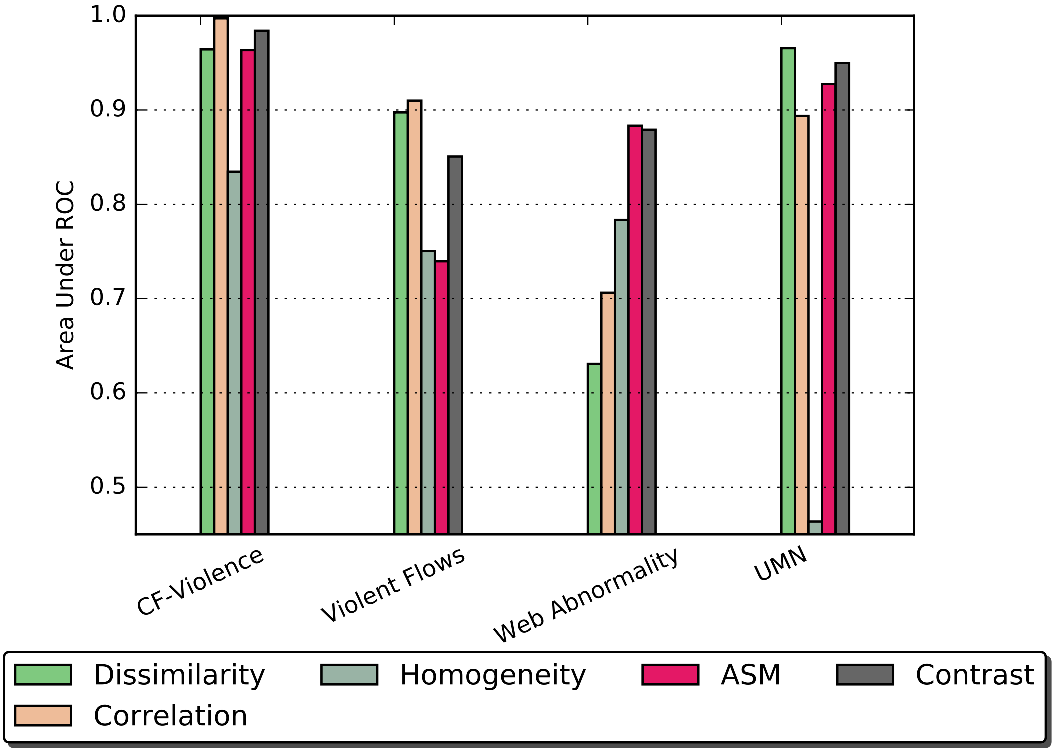


Figure 4.3: Classification performance achieved by each texture feature type.

Method	AUC
Proposed	0.9782
ViF [56]	0.80
OViF [40]	0.76
Fast Fight [49]	0.89

Table 4.2: CF-Violence classification score.

4.4.1 Pixel pair relationship

As stated in Section 4.3, a grey level co-occurrence matrix is generated by counting pixel pair occurrences given a relationship defined by parameters (θ, d). An investigation into the effects of (θ, d) was performed to determine whether or not

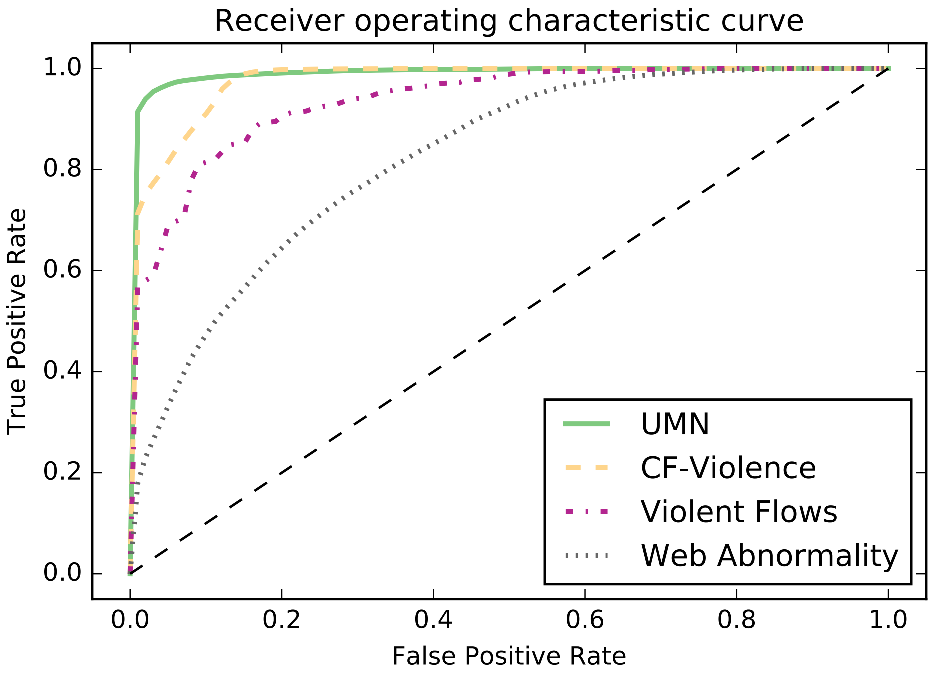


Figure 4.4: ROC curves for each tested dataset.

Method	AUC
Proposed	0.9956
Optical Flow [98]	0.84
SF [98]	0.96
MDT [90]	0.9965
Chaotic Invariants [152]	0.99
Biswas [6]	0.9838 (Average)

Table 4.3: UMN classification performance scores including state-of-the-art results.

a common value exists that offers good performance across different data types. The effects of these parameters were evaluated by performing multiple experiments with a range of values. The experiments use each combination that can be composed from one of three orientation configurations and one of five different

Method	AUC
Proposed	0.8218
SF [98]	0.73
Optical Flow [98]	0.66

Table 4.4: UCF Web Abnormality Crowd dataset classification performance scores including state-of-the-art results.

Method	Accuracy (±)	AUC
Proposed	86.03 ± 4.25%	0.9403
Fast Fight [49]	69.40 ± 5.0 %	0.7500
ViF [56]	81.30 ± 0.21 %	0.8500
OViF [40]	76.80 ± 3.90 %	0.8047
Holistic Features [93]	85.53 ± 0.17 %	–
MoSIFT [154]	83.42 ± 8.03 %	0.8751
MoSIFT (KDE / Sparse Coding) [154]	89.05 ± 3.26 %	0.9357

Table 4.5: Violent Flows dataset classification performance scores including state-of-the-art results.

distance values, this provides 15 different experiments. The first orientation configuration is a set of 8 orientation values spaced $\pi/4$ radians apart, the second set contains 4 orientation values spaced $\pi/2$ radians apart. The final orientation set contains a single value of 0. The five distance values are a sequence of integers that double in size starting from 1. The results of this experiment are shown in Figure 4.5. Across all tested datasets, a common trend is observed in that orientation has no significant impact on descriptive ability. A second experiment was conducted in which each cell was randomly rotated by either 0, 90, 180 or 360 degrees, No significant difference in classification performance were observed when using each of the three orientation configurations.

When analysing the results from the Web Abnormality, Violent Flows and CF-Violence datasets, a negative correlation between ROC and parameter d was reported; the relationship between ROC and parameter d is less uniform for the UMN dataset. It was hypothesised that this pattern occurs due to the distance between interacting entities within each video. For instance, the crowds structure depicted in the CF-Violence and Violent Flows dataset can be described as being densely populated. In densely populated crowds, pedestrians are typically in close proximity to one another. In contrast to this, the UMN dataset is sparsely populated and the distance between moving entities is much larger. In the close proximity scenario, a small distance value is better suited towards identifying meaningful relationships between two interacting entities, conversely, in a sparse scenario a small value for d may not be great enough to relate two distant entities, in which case a greater value of d should be chosen. Ultimately, it was determined that the parameter pair $(0, 1)$ provides the best performance across all datasets.

4.4.2 Cell Size

Described in Section 4.3 is a process where a frame is split into sub-regions, referred to as cells. These cells are described independently of each other to form a description of the local spatial region. A frame is then represented by aggregating the information from each cell into a single descriptor. This process is performed as a scene may consist of different local behaviours that will not be strongly represented when processing the entire scene as a single cell ($M = N = 1$). Presented in this section is a discussion of the relationship between cell parameters M and N, and classification performance.

The Violent Flows dataset sees maximal classification score when $N = M = 2$. When the scene decomposition becomes too fine $M = N > 8$ the classification performance drops, a similar trend occurs when analysing the Web Abnormality dataset. It is hypothesised that a larger cell size is more suitable for describing

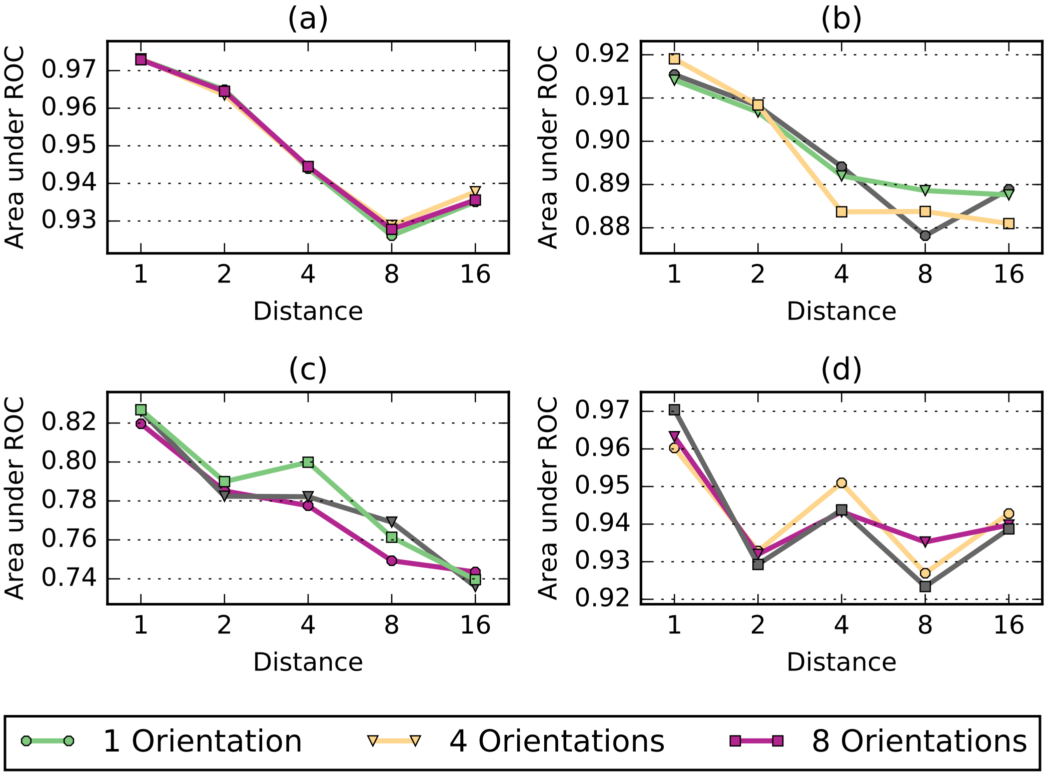


Figure 4.5: Graphs show the effects of pixel pair relationship parameters: a) CF-Violence, b) Violent Flows, c) UCF Web Abnormality, d) UMN Panic.

the global characteristics of behaviour in a dense crowd. Using a fine grid results in the description of small components such as individuals, in which case the characteristics encoding the effect of an individual on a crowd is less explicitly encoded as the local cell aggregation process used to form the global descriptor discards spatial locality of behaviour, and therefore any relationship between an action and its associated reaction is also discarded. Larger cell sizes will encapsulate multiple people and therefore describe the interaction, both action and reaction.

The CF-Violence dataset does not follow the same aforementioned trend. Instead a positive correlation between grid size and classification score is observed, suggesting that in this case, small individual components of a scene are sufficient

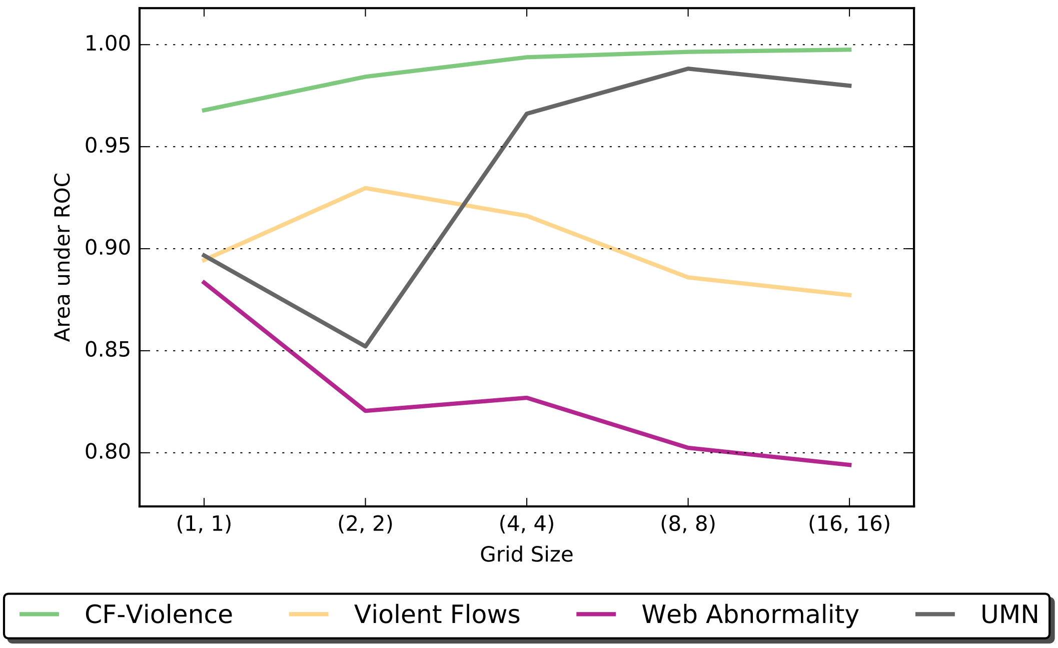


Figure 4.6: The effect on accuracy of using different values for M and N where $M = N$.

WE

enough to describe violent behaviour. This is reasonable when considering the contrast between normal and abnormal behaviour during the NTE, for instance, violent acts such as kicking or punching are vastly dissimilar to the typical types of normal behaviour, therefore smaller cells that encapsulate individual actions are still capable of encoding abnormal behaviour as its the action, not the interaction that matters. In contrast, the difference between the individual actions of people within the Violent Flows dataset during violent and non-violent scenes is less clear, and so the interactions are important, and as stated prior, encoding interactions require larger observation windows.

4.4.3 Window Length

In this subsection, an investigation into the effects of parameter n is conducted to determine if the description of crowd behaviour is best formulated using either short or long term temporal dynamics. The following values of n are used inform feature extraction: 6, 12, 24, 32, 64 and 128. Classification results show that all datasets favour larger window sizes (Figure 4.7), suggesting that the distinction between scenes of normality and abnormality is made more clear over long periods. Although each dataset has its preferential window length, it is important to note that short window sizes still offer reasonably good performance across all datasets. When transitioning from normal to abnormal behaviour, the amount of time required for the majority of the feature vector to be composed of information from abnormal behaviour will be greater the larger the observation window. Assuming that class transitions are not represented by the descriptor, the worst case for classifying abnormal behaviour will see a delay of at most n frames. Therefore shorter observation windows are more appropriate for use in a real-time system as it will allows for more instantaneous updates regarding the dynamics of the scene.

4.5 Conclusion

To summarise, GLCM texture features that are typically used in crowd density estimation were temporally encoded to create an effective method that describes crowd dynamics. It was shown that the proposed method is highly effective at discriminating between scenes of normal and abnormal behaviour. Additionally, the proposed approach is computationally cheap and operates in real-time. Analysis revealed that violent behaviour typically holds a less uniform rate of change over time when compared to other types of typical crowd behaviour, further analysis must be conducted to identify whether or not this property exists given alternat-

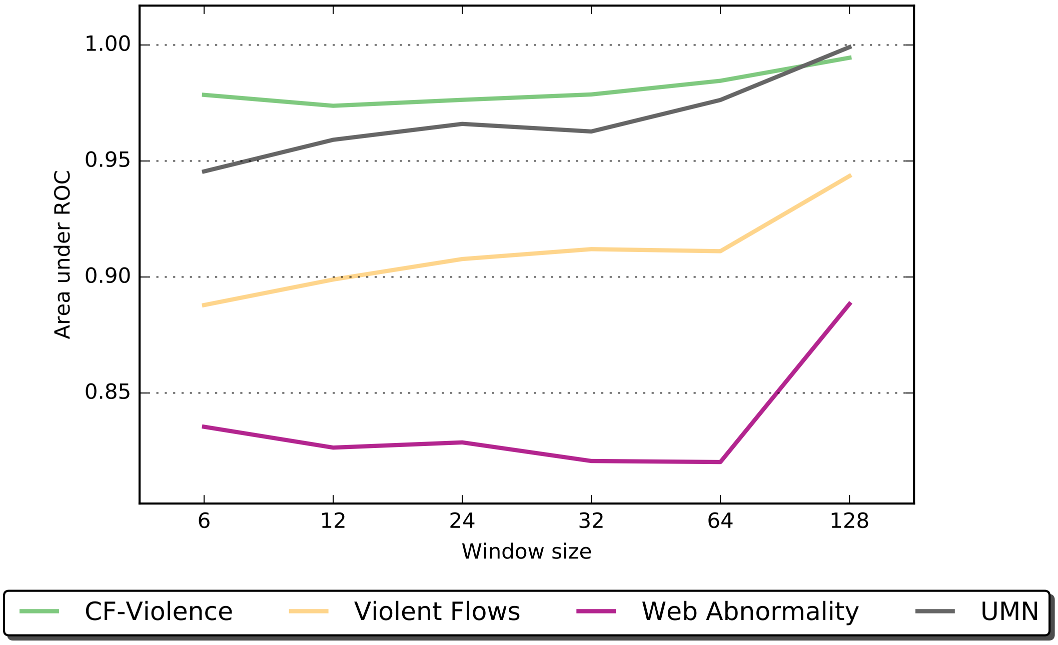


Figure 4.7: The effect on accuracy of using different window sizes n.

ive measurements of crowd behaviour. An in-depth evaluation of the parameter effects of the proposed method was conducted to provide insight for selecting suitable parameter values. Further research can be conducted to determine a method of adaptively choosing the optimal parameters given some data.

Chapter 5

Modelling One-on-One and Crowd Violence using Violent Interest Points

5.1 Introduction

The previous chapter concerns the subject of both *crowd* analysis and detecting violence that occurs in crowded environments. The approaches discussed in the previous chapter are not suited towards the analysis of scenes of violent behaviour that depict one-on-one violence. When presented with instances of one-on-one violence, the method presented in Chapter 4 performs poorly relative to existing research (Table 5.1). It was hypothesised that in a violent crowd, the violent behaviour is usually prominent and widely distributed due to the large population engaged in violence. In contrast, one-on-one violence can be described as being localised, a violent act between two individuals may not necessarily contribute much information to the overall scene. While two people fight, other actors in a scene may be behaving normally, continuing on their business. If the violent act does not contribute substantially to the scene, then the description of violence may be entangled and hidden within the description of other activities in the background. Although not directly intended, it is demonstrated that the approach presented in this chapter offers comparable performance with the state-of-the-art reported results for violent behaviour detection in crowds (Chapter 4).

The method presented within this chapter utilises an interest point detection methodology, in which actions associated with violence are detected and used to describe the nature of a scene. As informed by the literature, characteristics associated with violent behaviour in naturalistic environments are established and mathematically formulated. Literature informed feature design was partly motivated by poor performance reported after failed attempts at learning a useful feature representation using a deep learning architecture (Section 5.7).

Informed by the literature, the characteristics associated with violent behaviour are: *Motion Acceleration*, *Inverse Laminar Flow* and *Convergence*; these measures are computed from a set of dense motion trajectories. As the aforementioned characteristics are derived from dense motion trajectories, a matrix of values for each characteristic is returned. The size of each matrix is equal to the spatial dimensions of the source video. The resultant matrices represent the motion characteristics for objects at each pixel position of a video frame. Regions with higher values are more likely depict actions associated with violence. This information is used as a prior in an interest point detector scheme to produce a set of interest points based on actions that exhibit properties associated with violence. In this chapter, it is demonstrated that interest point sampling strategies based on violent characteristic priors produce a set of informative features. When encoded using a BoW scheme, the set of features from the interest point detector produce a more powerful description of a violent scene than features sampled using a regular grid structure.

Finally, an analysis of the underlying dynamics of violence reveals that the nature of violent behaviour can vary drastically, and that violence does not adhere to a singular definition with respect to the characteristics investigated.

5.2 Related Work

Both, Datta [26] and Deniz [29] produced methods that detect violence by identifying high motion acceleration, a property expected to belong to violent behaviour. Deniz notes that high acceleration often manifests itself as a *visual motion blur* which can be measured by identifying the shape of an ellipse in a radon transformed power spectrum composed using two consecutive images. Datta takes a more structured approach to solve the task of measuring person-on-person violence by first determining a person's silhouette, and subsequently their head, for tracking. The third derivative of motion, known as *Jerk*, is then incorporated in the composition of the Acceleration Measure Vector to describe violent motion. The drawback of this work is that it assumes a person's body is both visible and trackable, which in a city centre environment is not feasible due to occlusions caused by pedestrians in populated areas.

In contrast to person-on-person violence, Hassner *et al.* [56] looked at differentiating between violent and non-violent crowds. The authors introduced the Violent Flows dataset alongside the *Violent Flows* (ViF) feature vector that measures the average magnitude of dominant motions over time. Gao *et al.* [40] state that ViF does not capture behaviour whose orientation shifts while maintaining a constant velocity and demonstrated that combining their own Oriented ViF (OViF) and ViF increases classification ability.

Nievas *et al.* [104] extended the SIFT descriptor to work on optical flow data and created Motion Scale Invariant Feature Transform (MoSIFT). Nievas *et al.* [104] used a combination of MoSIFT and SIFT features to classify between violent and non-violent scenes that occur in ice hockey. Xu *et al.* [154] extended the work of Nievas and applied a kernel density estimation process to select the most representative features in a vector that were subsequently encoded using a sparse coding scheme. This allowed MoSIFT features to achieve excellent classification on person-on-person violence as well as on crowded data. Gracia *et al.* [49] argue that approaches such as these, although impressive, are too computationally costly to be practically implemented in the real-world so they propose a more efficient method of violence detection. Gracia *et al.* [49] perform adjacent frame difference and apply a fixed threshold to extract the largest blobs which are then described using measures of inter-blob distance and compactness.

Riberio *et al.* [113] introduce the Rotation Invariant Motion Coherence (RIMOC) feature that is based on the eigenvalues of second-order statistics extracted from a Histogram of Oriented Flows. A multi-scale structure is used to model spatiotemporal configurations of features. The authors assume that violent behaviour is unstructured and aim to distinguish this difference by analysing the likelihood of a feature belonging to a model of normality.

The work presented in this chapter is more akin to that of MoSIFT or STIP [77] but with a focus on using trajectory dynamics to identify spatio-temporal regions that exhibit non-linear, high acceleration interactions that are typical of violent behaviour. The key insight provided by our method is that utilising appropriate priors associated with violence results in more powerful scene description. The proposed approach is designed to function well on real-world surveillance data and is theoretically scale-invariant. Included in this chapter is a comparison of feature sampling strategies that demonstrates the benefit of using an interest points detection framework.

5.3 Proposed Method

Using a linear combination of three measures of motion trajectory, a response map that highlights spatiotemporal regions that are suggestive of violent behaviour can be generated. In this approach, motion trajectories are computed using a particle advection process that is widely used in the field of pedestrian analysis due to its robustness to small occlusions [98, 102, 162]. A multi-scale region selection process is used to extract spatiotemporal regions from the response map. Extracted regions are then described and for use in a classification model.

5.3.1 Particle Advection

The process of advection can be defined as the movement of an object through a medium guided by an underlying flow field; this process can be envisioned as a leaf in a river flowing downstream guided by the local flow of the water. In the context of our work, particle advection is performed by first generating a uniform set of particles that overlay the initial frame of a spatiotemporal volume. Each particle is advected using a Gaussian average of local optical flow vectors computed between successive frames. Particle advection is used to determine a set of motion trajectories T over τ frames for each pixel in a frame.

Inverse Laminar Flow

It has been observed that pedestrians walking through an environment would tend to exhibit laminar flow as they walked towards their destination providing their pathway is not obstructed [59, 156]. Pedestrians are not the only entities that exist in a city centre environment that exhibit this behaviour, with a vehicle being the most prominent example. After watching video footage, it was also observed that participants of a violent situation were often unstable in their movements as they attempted to perform violent gestures towards another person. Based on data observation and findings in the literature, it was hypothesized that inverse laminar flow is an indicator of potentially violent behaviour. For each trajectory T, two values representing the trajectory distance T_{dist} and trajectory displacement T_{dist} are computed. Trajectory distance T_{dist} is defined as the total distance travelled over τ frames. Trajectory displacement T_{dist} is defined as the absolute difference

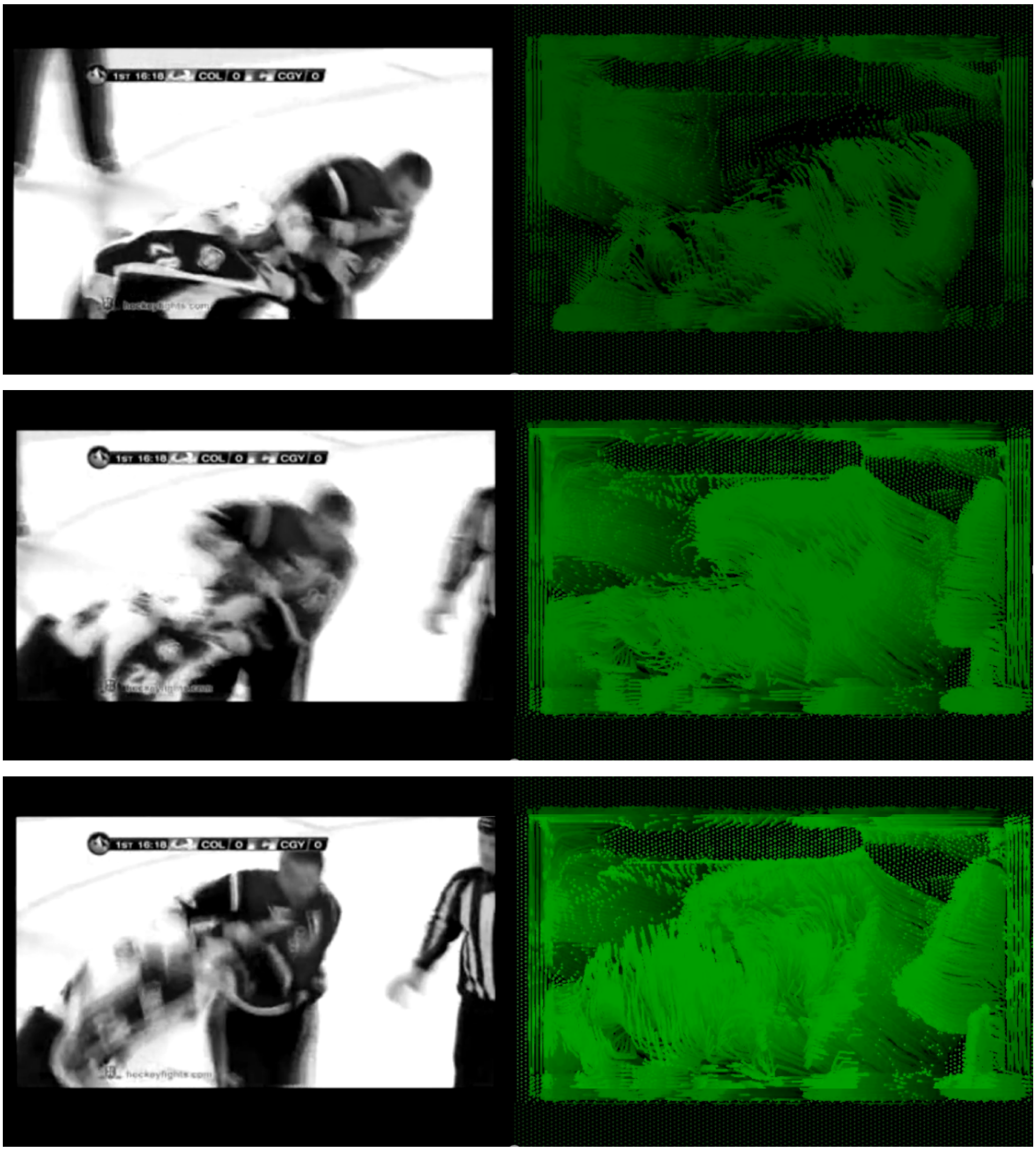


Figure 5.1: Visualisation of a set of dense trajectories extracted from a violent interaction.

between the trajectories initial starting position and its final position after τ frames. These two values are combined to form the inverse laminar flow response

R_{ilf} shown in Equation 5.1.

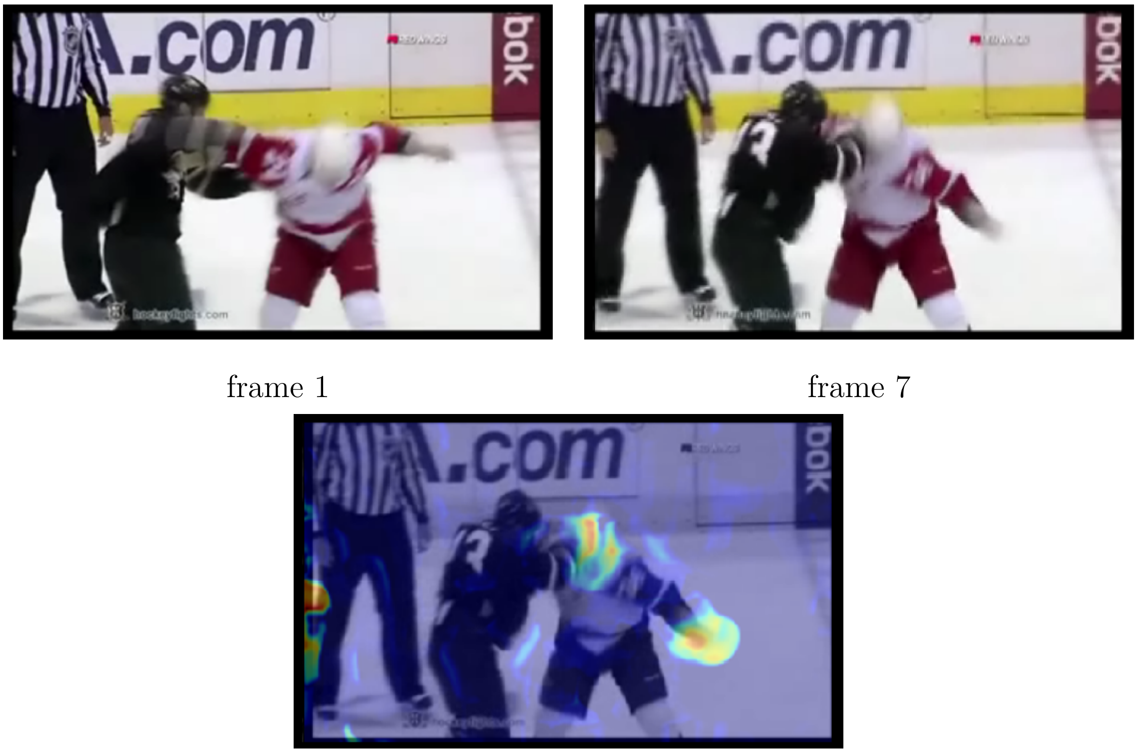
$$R_{ilf} = 1 - (T_{disp}/T_{dist}) \tag{5.1}$$

Acceleration Response

It has been observed that violent acts tend to show a greater increase in velocity when compared to normal behaviour. Typically, to injure a person during a fight, you must build up a substantial amount of force more quickly that your opponent, this process manifests as high acceleration. This was also one of the key principles behind the Maximum Warping Energy method of violence recognition [97]. To generate the acceleration response R_a, the pixel associated with the starting position of a trajectory is assigned the maximum acceleration value observed along the trajectory. The acceleration values are then divided by the maximum acceleration of all trajectories at a given point in time.

Motion Convergence

The location of an interaction between multiple entities can be identified by determining the point where they converge if they were to continue in their respective motion. For example, in a real-life scenario, this can manifest itself as two people approach one another, or to use a more extreme example, a fist moving towards a person's body. The local trajectory convergence response is used to describe whether or not particle trajectories interact. This is achieved by generating a 2D-histogram using each particle (x, y) position at the end of their trajectory. Histogram bins represents are representative of pixels in the source video. Convergence manifests within the 2D histogram as a bin whose value is greater than 1 as multiple particles share the same position in space, indicating that the particle trajectories converged onto the same position. The convergence



frame 1

frame 7

Figure 5.2: Example of the acceleration response output after analysing the acceleration properties of motion trajectories.

value of a trajectory is number of particles that share the same end position, as indicated by the 2D histogram. To form the response map R_c, a value is assigned to the pixel associated with the initial position of each particle equal to that of the convergence value obtained by the trajectory. During this process, to reduce noise and smooth the results, a Gaussian blur is applied to the 2D histogram after it is generated.

5.3.2 Response Map

Trajectory responses outlined in previous sections are linearly combined to form a single response map (Equation 5.2).

$$R = (w_1 R_{ilf} + w_2 R_c + w_3 R_a) \tag{5.2}$$

The limits of the combined response are $[0, 1]$, providing the sum of the three weights is equal to one. As scale invariance is necessary (see Chapter 3), the Difference of Gaussians (DoG) blob detection approach [81] is used to select the appropriate region size with which to represent an area in the response map. DoG is performed by applying different Gaussian filters to an image and subtracting the response. Applying the DoG process will generate a value for a given point in an image that representing the magnitude of difference between a focal area and its context; the difference is maximal when the gradient between context and focus is at its steepest. Applying the DoG process using Gaussian kernels of increasing size produces a set of responses that represent the magnitude of difference between focus and context at different scales. The output that results from subtracting two filtered images is multiplied by the σ value used to generate a Gaussian kernel, this is performed in order to normalise the DoG response and introduce scale invariance. The scale of an interest points is determined by identifying the maximum value across DoG responses of varying scales. In addition to this, non-maxima suppression is applied to reduce the number of detected interest regions. This process produces a set of locally contrasting regions and determines the optimal scale with which to represent an area of a video frame based on the combined response map.

When analysing appearance, the intensity of an object remains constant regardless of its spatial depth relative to the camera. Therefore, a DoG approach for detecting an object based on appearance should appropriately identify the object, albeit at a smaller scale. Motion data does not hold this property, the motion magnitude of an object drops as object depth increases, creating a situation where the distance objects cannot be reliably detected. Therefore, a scale normalisation step is introduced by dividing the value of the DoG filter response by the local maximum value of the combined response (Equation 5.2).

5.3.3 Feature Extraction

Motion velocity and acceleration data was obtained from trajectories that were initialized within an $N \times N$ area as dictated by the interest region and scale selection process. Motion velocity and acceleration are encoded by separating values temporally into three groups, each of which is used to generate a histogram that is then concatenated with each other in temporal order before being normalised using L_2 normalisation. Orientation was not encoded in the description of motion as features extracted from a scene where objects move vertically will not generalise well to describe a similar scene in which the dominant directionality of motion is horizontal; this example is common in CCTV surveillance systems. In addition to temporal motion descriptors, data from each trajectory response map that falls within the previously defined region is extracted. The aforementioned data is used to generate a histogram by placing values into β uniformly spaced bins before applying L_2 normalisation. Appearance is encoded using the Histogram of Oriented Gradients [24] approach.

5.4 Relative Response Importance

In this section, the underlying dynamics of violent and non-violent behaviour are compared by examining the distribution of each proposed trajectory response measurement outlined in Section 5.3. It was demonstrated that violent behaviour has different characteristics based on the context of the scene, and that violent footage is more closely characterised by non-linear, accelerating, convergent movements than non-violent footage. The processes described in Section 5.3 are applied to produce a set of interest points and their respective response values as dictated by the output produced post region/scale selection. The value of each point represents in some part the degree in which the corresponding underlying motion that generated it adheres to our definition of violence, which to reiterate, is stated to be a function of non-linear, accelerating, convergent movement. Given the lack of latent processes in the proposed method, the values of each component and their respective contribution to the final response value can be determined. Conducting this process post scale selection will provide an invariant measure of the importance of each element. Using the aforementioned process, it can be determined whether or not the three trajectory responses present different values when applied on violent and non-violent data.

Response values associated with each class, violent and non-violent, are used to generate a distribution of values for each measure and class. Kolmogorov-Smirnov and Kruskal-Wallis [74] tests are performed to determine whether or not the set of values for each measure extracted from violent and non-violent have a similar distribution and a similar median. Analysing the median will allow us to draw conclusions regarding the relative magnitudes of each response type given different classes of the same data. Figure 5.3 provides a visual representation of the median response per class where each distribution is composed of interests points detected using the parameters outlined in Section 5.5. In most cases, both the distribution and median of response values for each measure is significantly different between the violent and non-violent samples. The reported p-statistics for each test returned a value less than 0.05. There was no statistically significant difference between aggressive and violent behaviour in the NN-Violence dataset for the IFL attribute.

It is observed that each type of data used different aspects of motion to their advantage. Interest regions presented in the CF-Violence dataset shows an increased acceleration and convergence response for violent data compared to the non-violent counterpart; it is important not to interpret this as violent motions holding greater acceleration, but rather salient movements in a violent scene are locally more unique in their acceleration aspect. To further express this point, a non-violent action may present far greater acceleration values than those found in a violent scene. However, if these values are uniform across the scene, they yield a weak response in the presented approach, an action returns a high response when it is locally unique (locally maximal, maximally contrasting). A clear statement regarding the global rate of acceleration cannot be presented as the processes that aim to introduce scale invariance removes global information. This applies when interpreting all given response values using the outlined approach. Instead, statements presented in this section regard the relative importance of each attribute defined in Section 5.3 given the environment or context of an action.

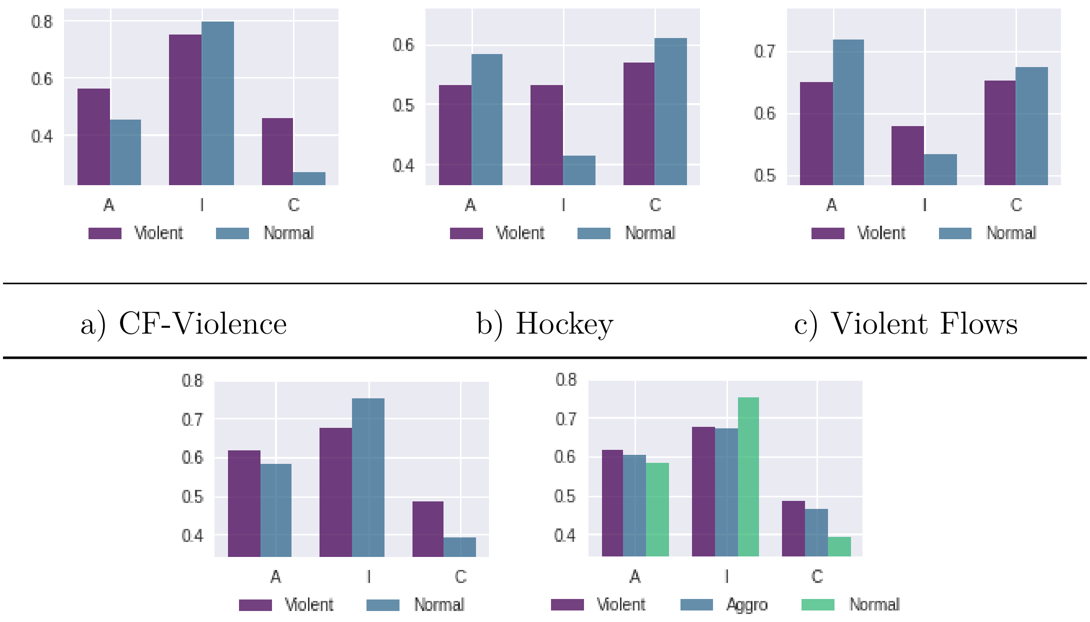
The data shows that in city centre environments, interest regions in a non-violent scene weakly utilise the convergence response; this suggests that in standard behaviour, few outlier interactions between pedestrians take place. This is inferred from both the NN-Violence and CF-Violence datasets (Figure 5.3). When examining the third class representing aggressive behaviour in the NN-Violence dataset, a middle ground is observed. Aggressive behaviour places a greater emphasis on convergence than non-violent actions, but to a lesser extent than violent behaviour; this expresses that the importance of analysing convergence is reflected in the escalation of violent behaviour. That is to say, as violence evolves, it is more important to focus on what objects are interacting and how these interactions take place.

For each dataset, the response values for each attribute has less variation for violent data when compared to non-violent data, suggesting that violent scenes utilise each attribute more evenly. In the non-violent case, at least one response type is vastly under-represented relative to the others This demonstrates that non-violent actions do not adhere to our definition of violence as strictly as behaviours presented in violent data.

Although violent samples adhere to our definition more closely that non-violence, the analysis in this section suggests that violence does not have an explicit static definition, but is instead modified by the environment and context of a scene. If we assume the characteristics of violence were constant, then we will expect to see similar relative response importance. Instead, it was found that violence captured by different CCTV systems are similar in characteristics, but different to violence associated with hockey and crowds. This understanding is corroborated by results presented in Section 5.6 where methods designed to analyse violent crowds perform relatively poorly when analysing CCTV data. It is important to consider the variation in violent behaviour when creating a general solution to the violence detection problem.

5.5 Experimental Setup

The ability of an algorithm to detect violent behaviour using a binary classification approach to separate feature vectors into one of two classes, violent and non-violent, is evaluated in this section. Results are reported using overall classification accuracy and receiver operating characteristic scores. When testing using the NN-Violence and CF-Violence datasets, we performed K-fold cross-validation with K equal to five and evaluated the mean score across all folds. Classification was performed using the Scikit-learn implementation of the Linear SVM classifier. Testing methodologies outlined in their respective seminal papers are used when evaluating violence detection algorithms using the Hockey [104] and Violent Flows [56].



a) CF-Violence b) Hockey c) Violent Flows

d) NN-Violence e) NN-Violence, Three Class

Figure 5.3: Median response separated by action class and individual response type where I, A and C represent Inverse Laminar Flow, Acceleration and Convergent response respectively.

Feature extraction at each frame position of a video was performed using trajectories of length τ. K-means clustering was to create a visual vocabulary for feature encoding based on visual word occurrences. A vocabulary size of 500 was empirically chosen as it resulted in high classification performance without requiring a large amount of time to generate the vocabulary, allowing for faster testing. Vocabulary size parameter optimisation may further increase performance. Each frame in a video is represented using an L_2 normalised histogram of word occurrences based on the features that occurred in the past W_t frames. A small parameter for W_t may not capture the co-occurrence of features in time and subsequently miss crucial relationships between features. In the case of the CF-Violence and NN-Violence datasets the value of W_t is set to twice their respective frame-rate. One key variable in the proposed approach is the number of frames a particle is advected along, otherwise referred to as trajectory length τ throughout this chapter. Given that short-term motion typically characterises violent actions then a small value for τ should be adequate to capture the properties of violent behaviour. $\tau = 8$ for all experiments as it allowed for real-time feature extraction. To reduce computation time, videos are resized so that their spatial dimensions are 160×120.

The weights for each response map (w_1, w_2, w_3) were all assigned a values of $\frac{1}{3}$ so that each response type is treated equally and that final response R is bounded between $[0, 1]$. A response threshold is applied to suppress weak features. The threshold value is equal to one twentieth the maximum weight for a given frame. A proportional threshold avoids situations where no features are detected, this can occur when all response values fall below a static threshold. A static threshold can result in zero detected features, and without features, we have no information to train a classifier. The proposed approach, naively implemented in C++ and CUDA and using the previously defined parameters operates at ≈ 55 frames per second when working on a Nvidia 760 GTX GPU, and i7-4790 CPU at 3.60Ghz.

5.6 Results

Results reported by the proposed method are compared with results obtained using the Violent Flows, Oriented Violent Flows [56], Fast Fight detection [49] and MoSIFT [104] methods of violence detection. Additionally, a comparison is made with results obtained by classifying features extracted using the C3D deep learning architecture [136] trained on the Sports 1M datasets [70]. Experimentation has shown that the proposed approach, in general, offers good all round classification performance when trying to determine whether or not a scene displays violent behaviour. Table 5.1 indicates that the proposed approach achieves comparable performance against existing methods when classifying one-on-one violence as shown in the Hockey violence dataset. The application of the proposed method on the Violent Flows dataset, like before, attains comparable results with existing methods.The application of the ViF and OViF descriptors on the Violent Flows dataset as presented in their respective papers utilise global feature extraction; global feature extraction encodes global spatial structure, generating a representation that is not robust to both rotation and translation. It is arguably acceptable to apply a global description method to the Violent Flows dataset as the events depicted are centrally focused. Events presented in the other datasets are not centrally focused; therefore local feature description is preferable over global approaches when analysing surveillance footage as the unfolding events are not guaranteed to be the central focus of the camera. The application of ViF and OViF on the NN-Violence and CF-Violence datasets utilise the same approach used to test the Hockey violence dataset as explained in their respective papers. Briefly stated, the authors apply their feature description to a set of interest regions identified using a Space-Time Interest Point detector [77]; a Bag of Words model is produced, and classification is performed based on feature occurrence within a spatiotemporal volume. It is demonstrated that the proposed approach outperforms all other tested methods when applied to the city centre surveillance datasets, NN-Violence and CF-Violence (Figures 5.4 and 5.5).

For comparison, the required computation time for each tested method on the violent flows dataset is presented in Table 5.3. The Violent Flows dataset was chosen as a point of reference for two reasons. The dataset is publicly available which allows for direct comparison with potential future research. Secondly, the Violent Flows dataset depicts dense population which results in more detected interest points than any other dataset, creating a worst-case runtime scenario for interest point detection based approaches.

A second experiment using the CF-Violence dataset was conducted in which interest regions found using the proposed approach were described using the ViF

Method	Classifier	Hockey	Violent Flows	CF-Violence	NN-Violence
Proposed	Linear SVM	0.97	**0.94**	0.96	**0.76**
Chapter 4	Random Forest	0.87	**0.94**	**0.98**	**0.74**
Fast Fight	Random Forest	0.90	0.75	0.89	0.59
ViF	Linear SVM	0.88	0.88	0.80	0.62
OViF	Linear SVM	0.90	0.81	0.76	0.62
MoSIFT	Linear SVM	**0.99**	0.88	-	-
C3D	Linear SVM	**0.99**	0.90	**0.96**	0.56

Table 5.1: Receiver Operating Characteristic scores for each dataset.

Method	Classifier	Hockey	Violent Flows	CF-Violence	NN-Violence
Proposed	Linear SVM	91.1 ± 1.9	87.3 ± 3.1	94.4 ± 5.0	74.9 ± 3.4
Fast Fight	Random Forest	82.4 ± 0.6	69.4 ± 5.0	85.4 ± 5.0	71.5 ± 10.3
ViF	Linear SVM	81.6 ± 0.2	81.2 ± 1.8	64.6 ± 6.4	66.1 ± 6.0
OViF	Linear SVM	84.2 ± 3.3	76.8 ± 3.9	59.2 ± 8.6	66.3 ± 6.6
MoSIFT	Linear SVM	96.7 ± 0.7	83.4 ± 8.0	-	-
C3D	Linear SVM	94.9 ± 1.54	83.7 ± 4.1	98.0 ± 3.1	72.6 ± 18.5

Table 5.2: Reported accuracy and standard deviation for each dataset.

and OViF descriptors. Classification results indicate an increase in ROC score when compared to the same description of areas identified by STIP (Table 5.4); this suggests that the proposed method of region extraction identifies more relevant regions of a scene when analysing violent behaviour.

5.6.1 Feature Sampling Strategies

Existing research suggests that dense feature sampling using a grid structure outperforms salient region detection frameworks for both image [107] and video [124,

Method	FPS	time per frame (ms)
Proposed	44.0	0.0227
Chapter 4	76.9	0.0131
Fast Fight	660	0.0015
ViF	40.8	0.0245
OViF	19.5	0.0512
C3D (Nvidia GTX 760)	2.7	0.3773
C3D (Nvidia Titan X)	30.1	0.0332

Table 5.3: Average operating-time for each method using the Violent Flows dataset.

Method	Detector	Classifier	ROC	Accuracy
ViF	Ours	SVM	0.88	76.8 ± 4.6
OViF	Ours	SVM	0.82	65.7 ± 9.0
ViF	STIP	SVM	0.80	64.6 ± 6.4
OViF	STIP	SVM	0.76	59.2 ± 8.6

Table 5.4: Results obtained when utilising the ViF and OViF descriptors to describe regions identified using the proposed solution and similar regions identified using STIP.

144] based classification tasks. The general principle is a dense set of features can be efficiently sampled using a set of overlapping grid structures. Wang *et al.* [144] conclude that such an approach provides greater classification ability when processing real-world data. Similar findings by Shi *et al.* [124] informed the design of the RIMOC method of violent behaviour detection [113]. In this section, the effects of feature density and differences between feature sampling strategies are investigated and discussed.

The density of features identified using the proposed approach is determined by

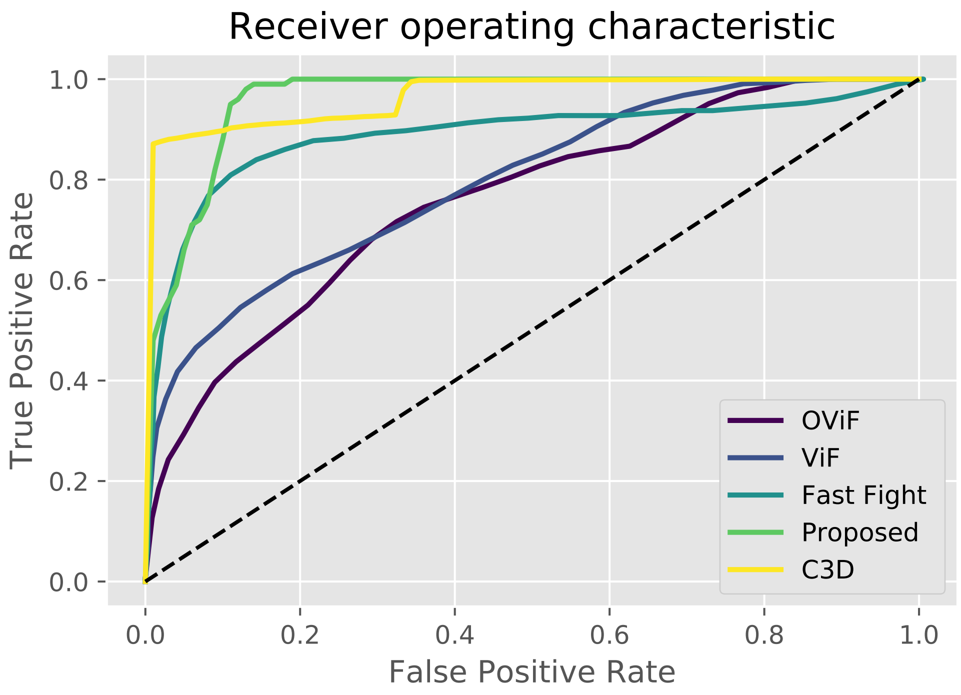


Figure 5.4: Receiver operating characteristic curve for each method tested on the CF-Violence dataset.

a threshold value which filters out weak responses to produce a set of interest points. The threshold is computed by multiplying the maximum response R for a given frame by some real number t which lies in the range $[0, 1]$. The quantity of features is directly related to the computational cost of the method, the fewer regions that must be described, the faster the system. The ideal threshold for a computationally efficient system should maximise classification performance and minimise feature sample count. Figure 5.6 plots the ROC performance metric against the average feature count per-frame for each dataset given the values of $0.05, 0.1, 0.2, 0.3$ for t. For each dataset, the relationship between the average number of features per-frame and classification performance is observed as being positive. These results corroborate existing research on image classification tasks [107] that shows that the more samples used, the greater the classification ability. The spatial distribution of identified regions may have an affect

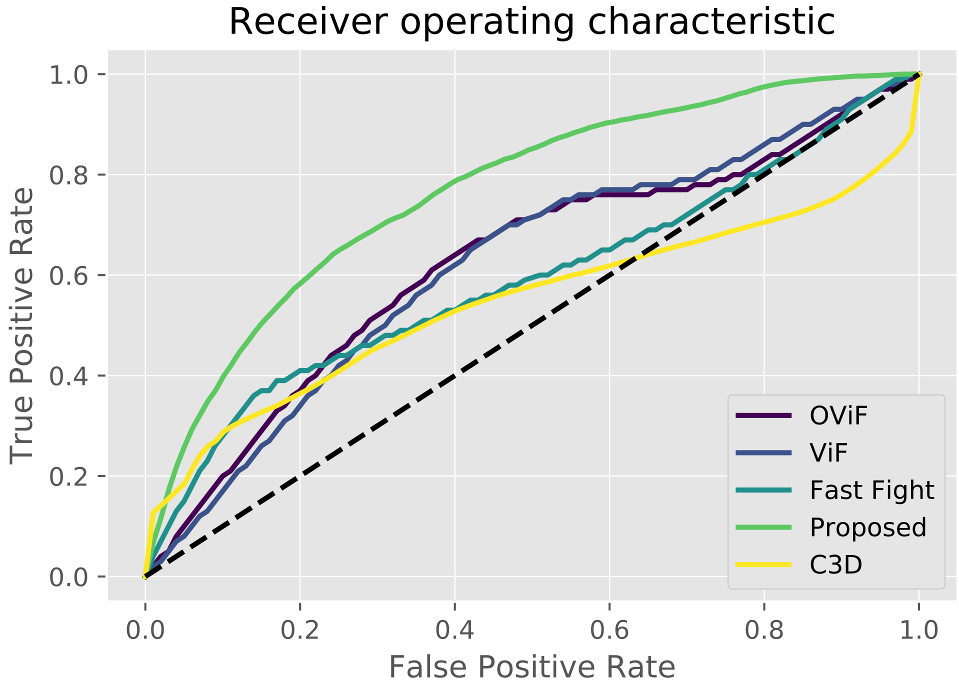


Figure 5.5: Receiver operating characteristic curve for each method tested on the NN-violence dataset.

on the classification performance. A measure that quantifies the proportion of a frame that is described by the interest points is used to analyse the spatial distribution of the sampled features. This is achieved by counting the number of unique pixels that fall within an interest region and dividing by the number of pixels in an image. Figure 5.7 depicts the relationship between feature count and the average proportion of a frame that is described. As the feature count increases, the proportion of the frame that is described tends towards one. The results show that the more of the frame that is described, the greater the classification score (Figure 5.8). It is understood that the positive relationship observed demonstrates that dense sampling increases the scope of the description, capturing more actions and background information provides contextual information. It is important to investigate whether the distribution of features sampled using the proposed approach provides any benefits over a grid-based sampling approach. The dense-grid sampling method described by Wang *et al.* [144] is used. In the cases of the NN-Violence, CF-Violence and Violent Flows datasets, it was observed that the proposed approach offers greater classification ability (Figure 5.9). The dense grid-based sampling approach is guaranteed to describe regions of a frame that can be considered noise. It is hypothesised that a dense grid-based sampling approach gives equal power to noisy, background information than it does the meaningful information. In comparison, the proposed approach is not guaranteed to sample data that is considered noise, and that it is also capable of uneven spatial sampling in which an informative region in a frame may be sampled more frequently than other areas, giving less power to features that describe noisy regions. There is no difference between the two sampling methods when analysing the Hockey violence dataset.

There is a strong positive correlation between classification performance, and screen coverage. The more of the screen that is described, the greater the score.

5.6.2 One Class Learning

When evaluating whether a new sample belongs to either the violent or non-violent class, there exists a fundamental assumption that the data you feed into a system belongs to one of the two learnt classes. In the CCTV datasets, CF-Violence and NN-Violence, there are a small number of violent samples. Therefore, there is a possibility that there exists an unseen violent sample that breaches the fundamental assumption by not belonging to either of the learnt distributions for non-violence and violence. Work presented in Section 5.4 demonstrates that violent behaviour in different contexts can have different characteristics, giving evidence towards the possibility that there exists a sample that breaches the fundamental assumption of classification. In this hypothetical case, the classifier would be expected to perform similarly to a random classification process. As discussed in Section 5.6.1, dense feature sampling that covers a large proportion

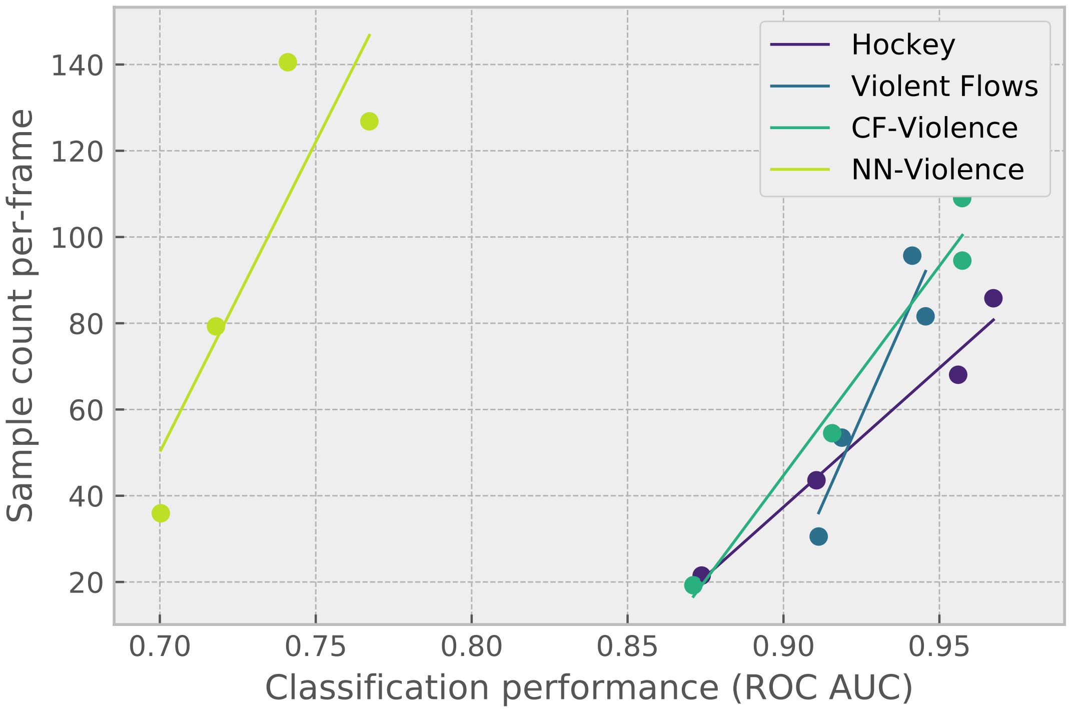


Figure 5.6: Relationship between sample count per-frame and classification performance (ROC).

of the background results in greater performance. This suggests that scenes of violence and non-violence are best described with context. Assuming this is true, if we change the context of a violent scene, but keep the violence identical, then it is likely that we would misclassify a violent sample. Given this theoretical situation, it is difficult to evaluate whether or not our models generalise to sufficient degree to avoid this issue. Although unlikely, in the worst case scenario, all violence that exists in the datasets is unrepresentative of the true distribution of violent behaviour. In an attempt to quantify the potential performance of the classifier in such a situation, one-class learning methodology is adopted. A model of normality is created by fitting using only the data that represents non-violence. Using this model of normality, the likelihood that a new sample belongs to a learnt non-violent model can be determined. If the likelihood is low, then the sampled is considered depicting *violence*, or more generally, as depicting

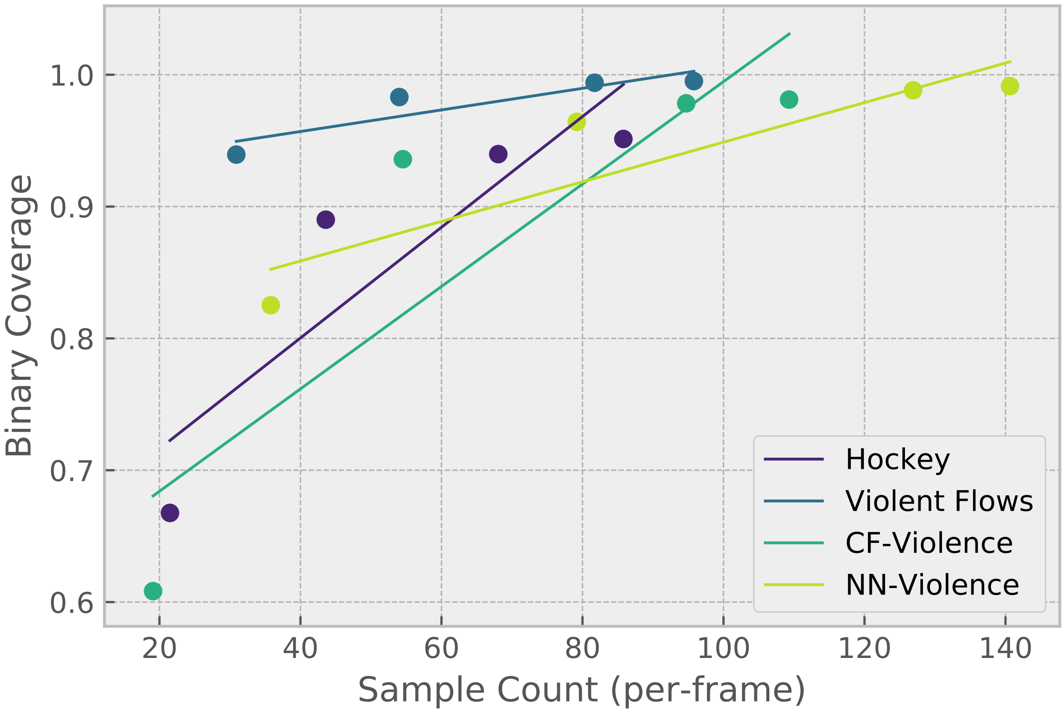


Figure 5.7: Relationship between the proportion of the frame described and feature count per-frame.

abnormal behaviour.

To perform one-class classification, non-violent data is used to fit a Gaussian Mixture Model using an Expectation Maximisation (EM) algorithm. To determine the number of mixture components in our model, multiple models using various component values are trained, and the model that minimises the Akaike Information Criterion (AIC) is selected. The potential number of components in each model are $[10, 20, 40, 80, 160]$. A similar methodology has been used to detect abnormal behaviour in an underground train station [137]. In comparison with the binary learnt model, a drop in classification performance as measured using the ROC metric is observed (Figure 5.10). CCTV footage captures behaviour that takes place within a mostly consistent and static environment. The effects of the environment, both physical and social, influence the behaviour of pedestrians to

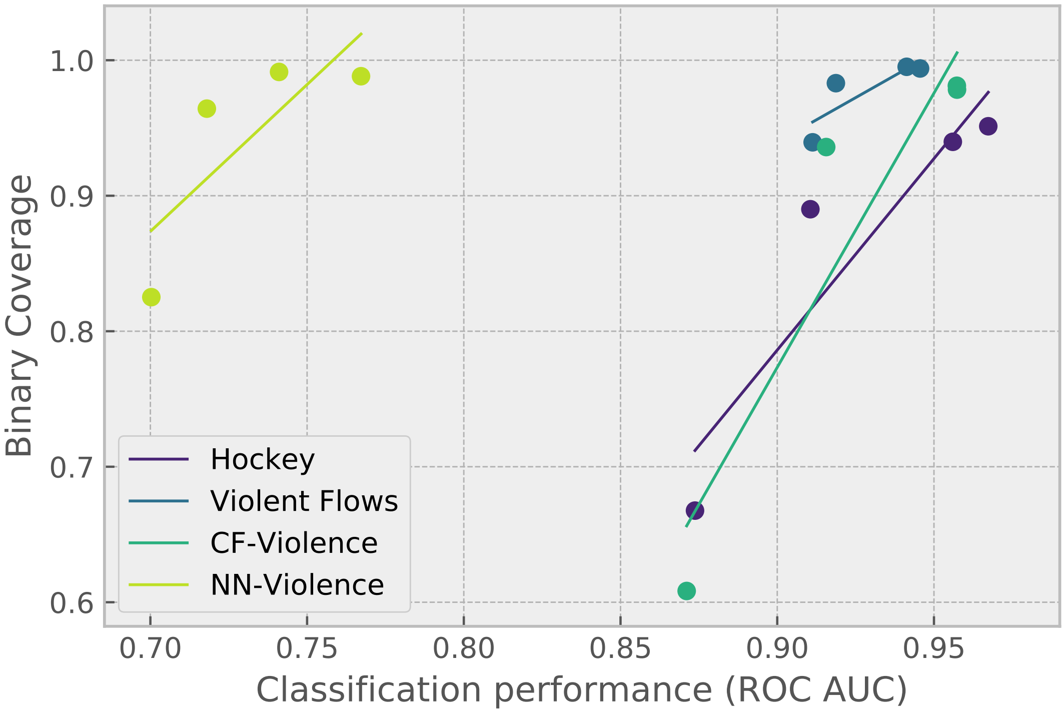


Figure 5.8: Relationship between the proportion of the frame described and classification performance.

produce a common mode of behaviour. Due to the influence of the environment, data in the CCTV datasets that is non-violent is similar. In contrast, the Violent Flows dataset contains footage obtained from a wider range of unique environments, increasing the variation in types of normal behaviour. Due to the difference in variation of behaviour, it is believed that the one-class learning methodology more easily learns a model of normality for the CCTV datasets than it does for the Violent Flows dataset as there are fewer outlier behaviours. This is reflected in the difference in ROC measure between one-class and two-class learning, the difference is greatest for the Violent Flows dataset (Figure 5.10).

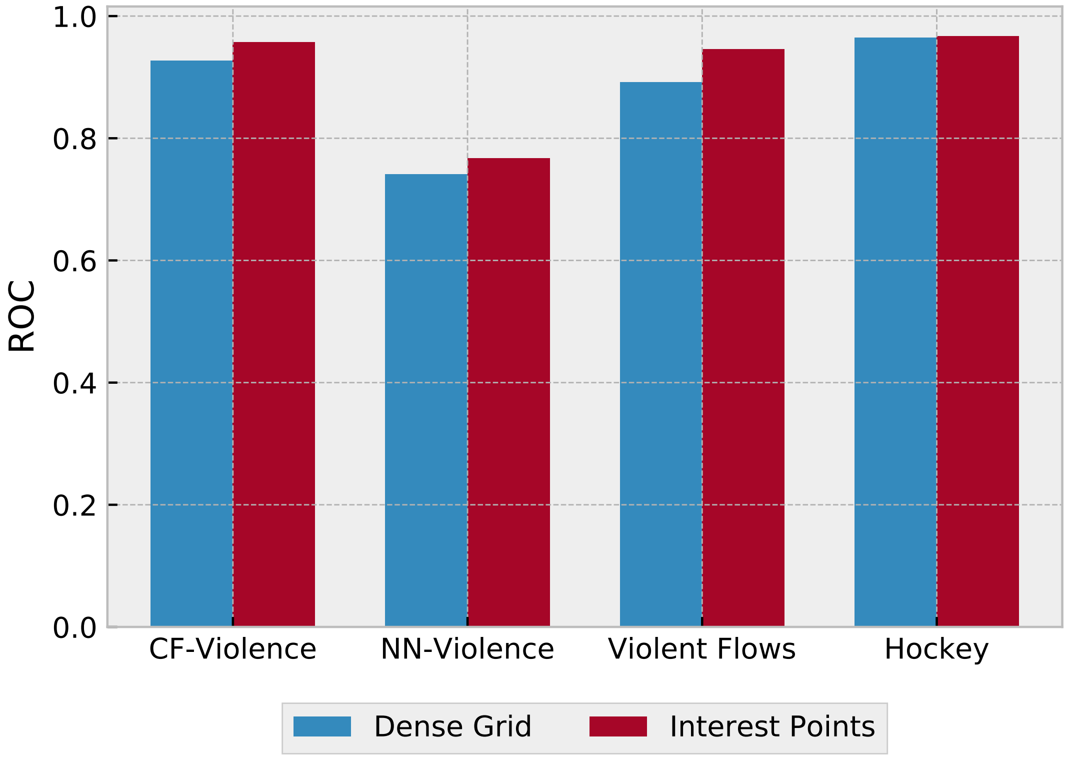


Figure 5.9: Comparing dense grid and interest point sampling methods. This bar plot demonstrates the difference in ROC classification performance achieved by a dense grid sampling approach and the proposed method.

5.6.3 Model Transfer

As discussed in Section 5.4, the kinetic definition of violence is not singular, but appears to be modulated by other factors such as environment, context or capture mechanisms. This observation is further tested by attempting to transfer knowledge learnt from one dataset in order to classify another. The classification results presented in Table 5.5 demonstrate the classification performance achieved when training a model using one dataset, and testing using another. Poor data transferability is reported as indicated by the relatively poor classification results.

It appears that the CF-Violence dataset proves to be a poor candidate for model

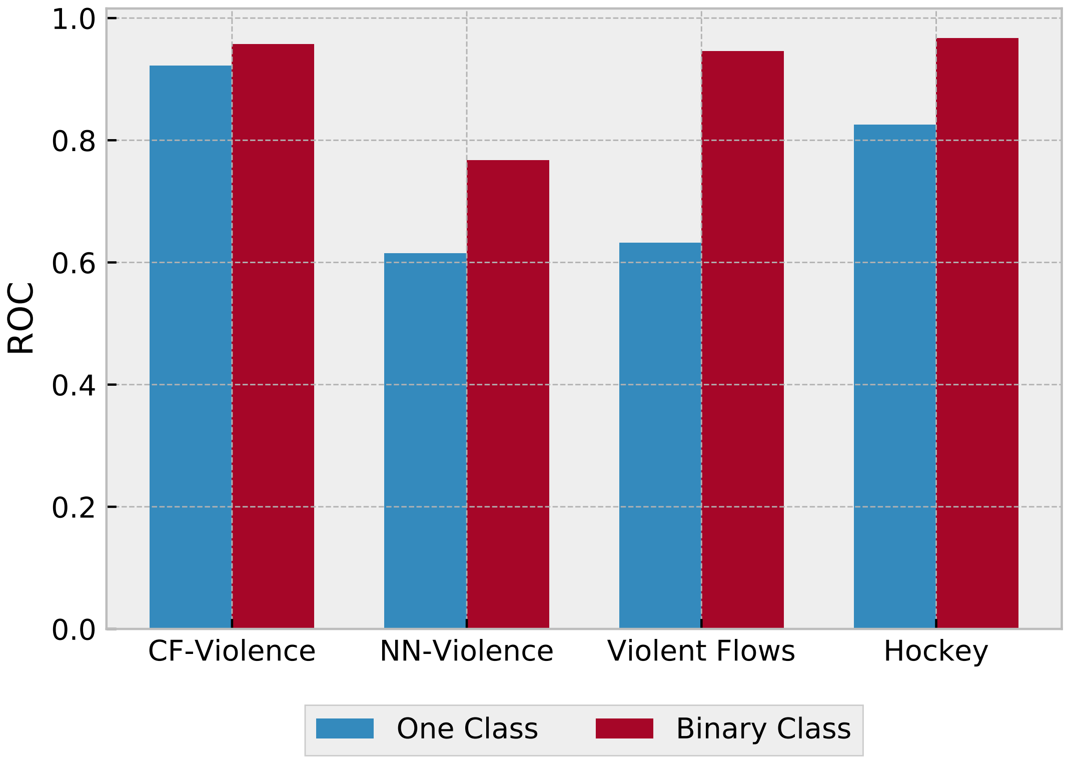


Figure 5.10: Comparison between one class and two class learning methods. This bar plot demonstrates the difference in ROC classification performance achieved by a one class, and two class learning scheme.

training, potentially due to the low violent sample count. The low sample count provides a small subset of potential violence, resulting in a model that does not generalise well. Models trained on the Hockey and Violent Flows datasets capture characteristics that allow for non-random classification performance. However, the results obtained are relatively distant from the maximum classification score as presented in Table 5.1. This may suggest that the descriptor fails to capture many important aspect of violence that generalise well. An alternative hypothesis is that the datasets lack common aspects due to large differences in performed actions and environments in each dataset. Due to hardware limitations, experiments involving the NN-Violence dataset failed to execute.

Training	Testing	AUROC
CF-Violence	Hockey	0.51
CF-Violence	Violent Flows	0.55
Hockey	CF-Violence	0.76
Hockey	Violent Flows	0.72
Violent Flows	CF-Violences	0.83
Violent Flows	Hockey	0.75

Table 5.5: Reported AUROC by training a linear SVM classifier using one dataset and testing using another.

5.7 Deep Learning: Training from scratch

The results achieved by the C3D network as presented in Section 5.6 utilised a pre-trained model. For a given input video, a set of features were extracted based on the information present in the pre-trained model. The model was pre-trained using the Sports1M dataset, chosen for the inclusion of fast actions that are similar to those seen in violence. Following is a brief overview of the results obtained when attempting to fit a C3D model using only data contained within the violent behaviour datasets exclusively. Understanding how well the C3D model fits the data requires looking at the reported model accuracy and loss values, and analysing how they change each time the model weights are updated. Results will be interpreted with respect to the training and validation set. The model is fit and evaluated using the training set to produce the training loss and training accuracy. The validation loss and accuracy is obtained by evaluating the model using unseen data.

As presented in Figures 5.11, 5.12 , 5.13 and 5.14, the accuracy values reported by the training set for each dataset appears to converge and stabilise after a few epochs. The training accuracy is low for each dataset with the CF-Violence reporting the highest value of 75% accuracy; the remaining datasets report a training accuracy close to 50%. The validation accuracy fails to stabilise and is lower than the training accuracy with the exception of the Hockey dataset.

In each case, the loss value appears to converge on a value after a few epochs. This may suggest that the model weight optimisation process is stuck in a local minima, failing to update the model weights in any significant capacity. In response to this, the learning rate of the Stochastic Gradient Descent (SGD) optimiser was increased from 0.001 to 0.01 after 30 epochs, to no great effect.

The reported results indicate that training a model from scratch to learn a feature representation for violent behaviour detection is difficult. The literature seems to agree with this statement given that existing work that concerns deep learning for violent behaviour detection utilise pre-trained networks and/or transfer learning strategies to achieve good results for violent behaviour detection [94, 32, 143, 99]. It is hypothesised that the amount of available data for the violent detection task is too low to learn a useful representation. For instance, the Violent Flows dataset contains 123 samples per-class, far below the rule of thumb of 5000 per-class [45].

5.8 Conclusion

In this chapter, three measures of dense motion trajectories have been proposed that when combined produce a response map that highlights regions within a spatio-temporal volume that contain behaviours associated with violence. The violence response is normalised and sampled using a Difference of Gaussians approach to achieve scale invariance. It has been experimentally demonstrated that the proposed approach achieves state-of-the-art performance across a wide variety of violence detection datasets. Furthermore, two CCTV surveillance datasets were used to evaluate the ability of multiple violent behaviour detection methods at detecting violent behaviour captured by real-world surveillance systems.

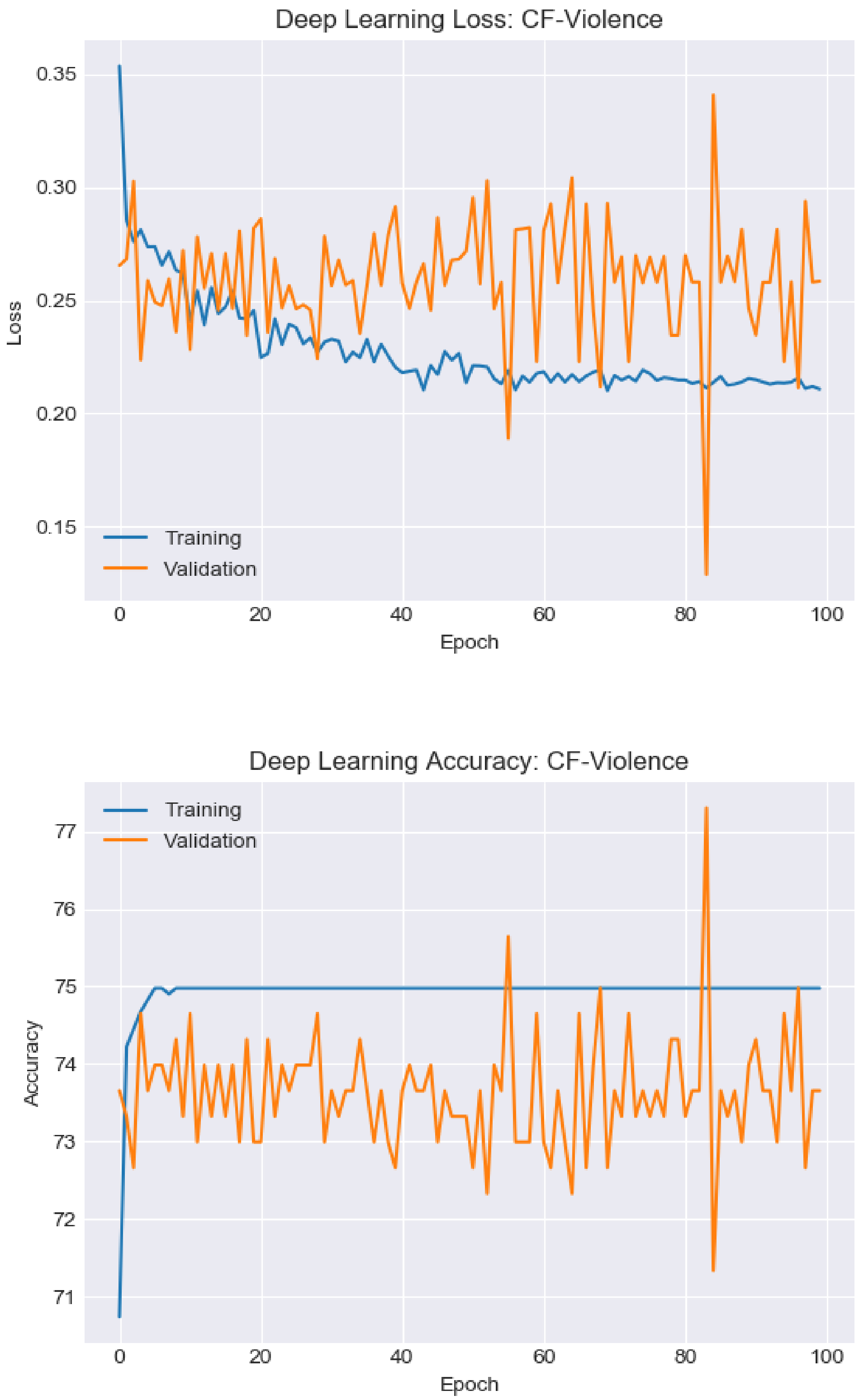


Figure 5.11: Training and validation accuracy/loss when fitting a C3D model using the CF-Violence dataset.

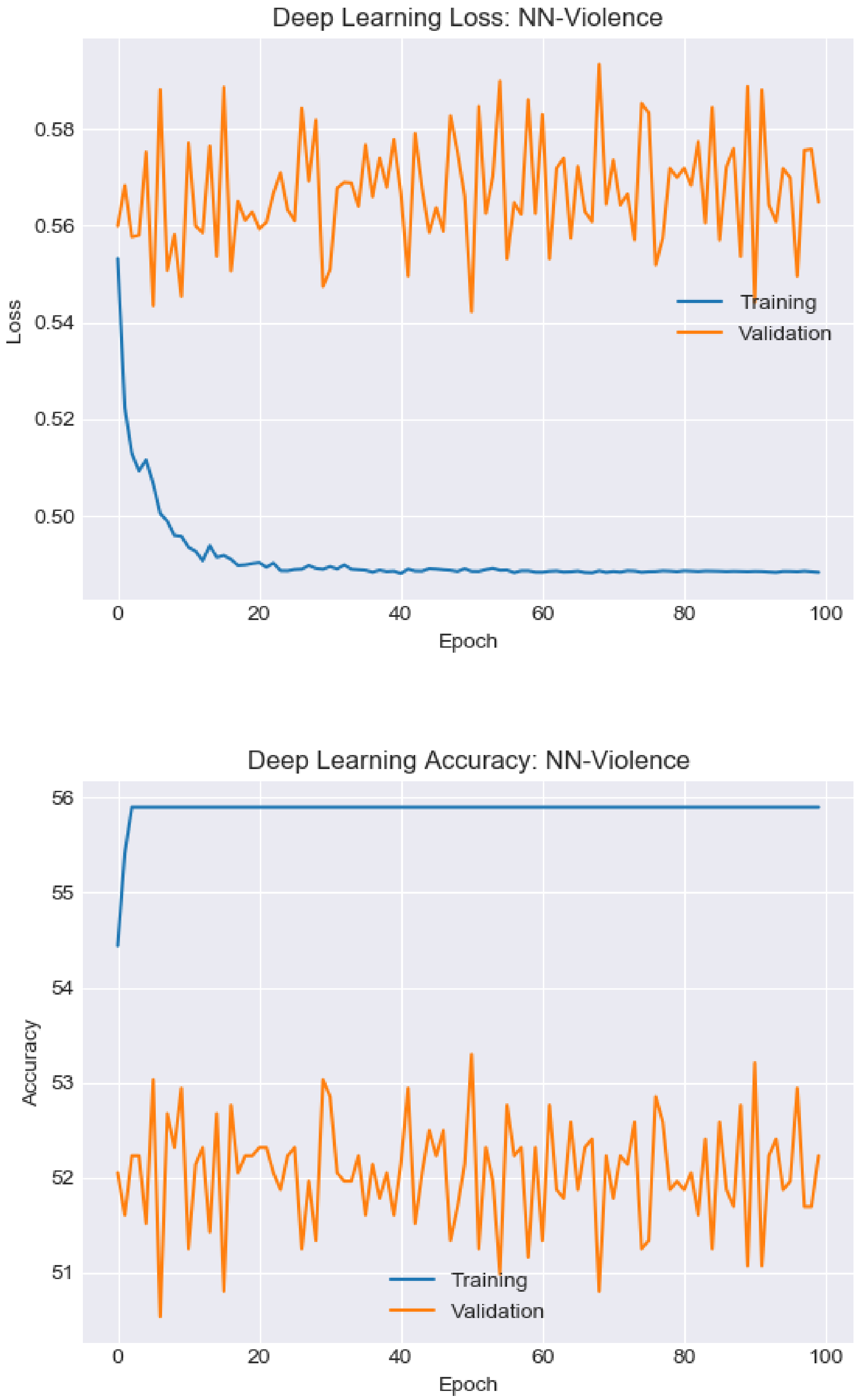


Figure 5.12: Training and validation accuracy/loss when fitting a C3D model using the NN-Violence dataset.

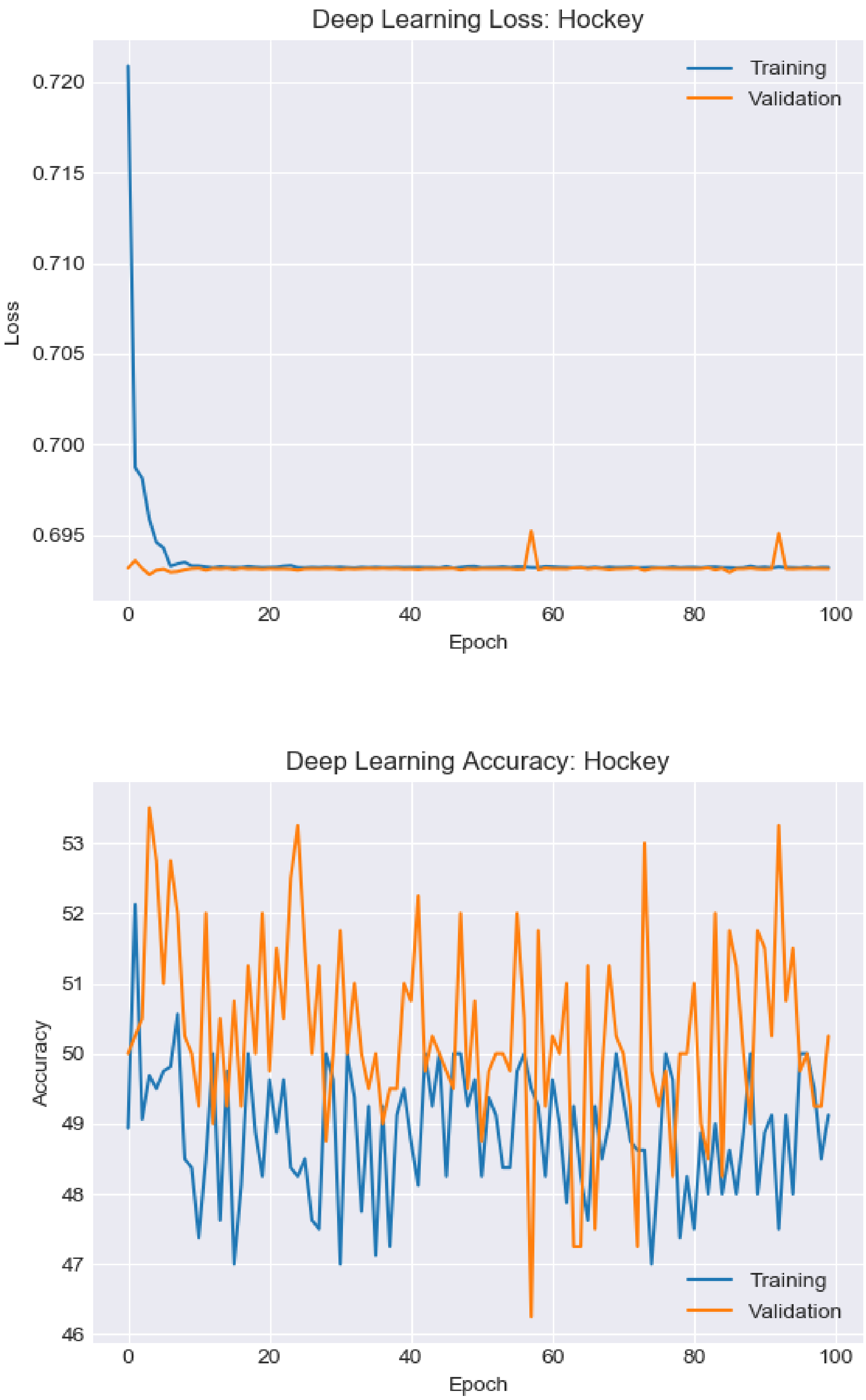


Figure 5.13: Training and validation accuracy/loss when fitting a C3D model using the Hockey dataset.

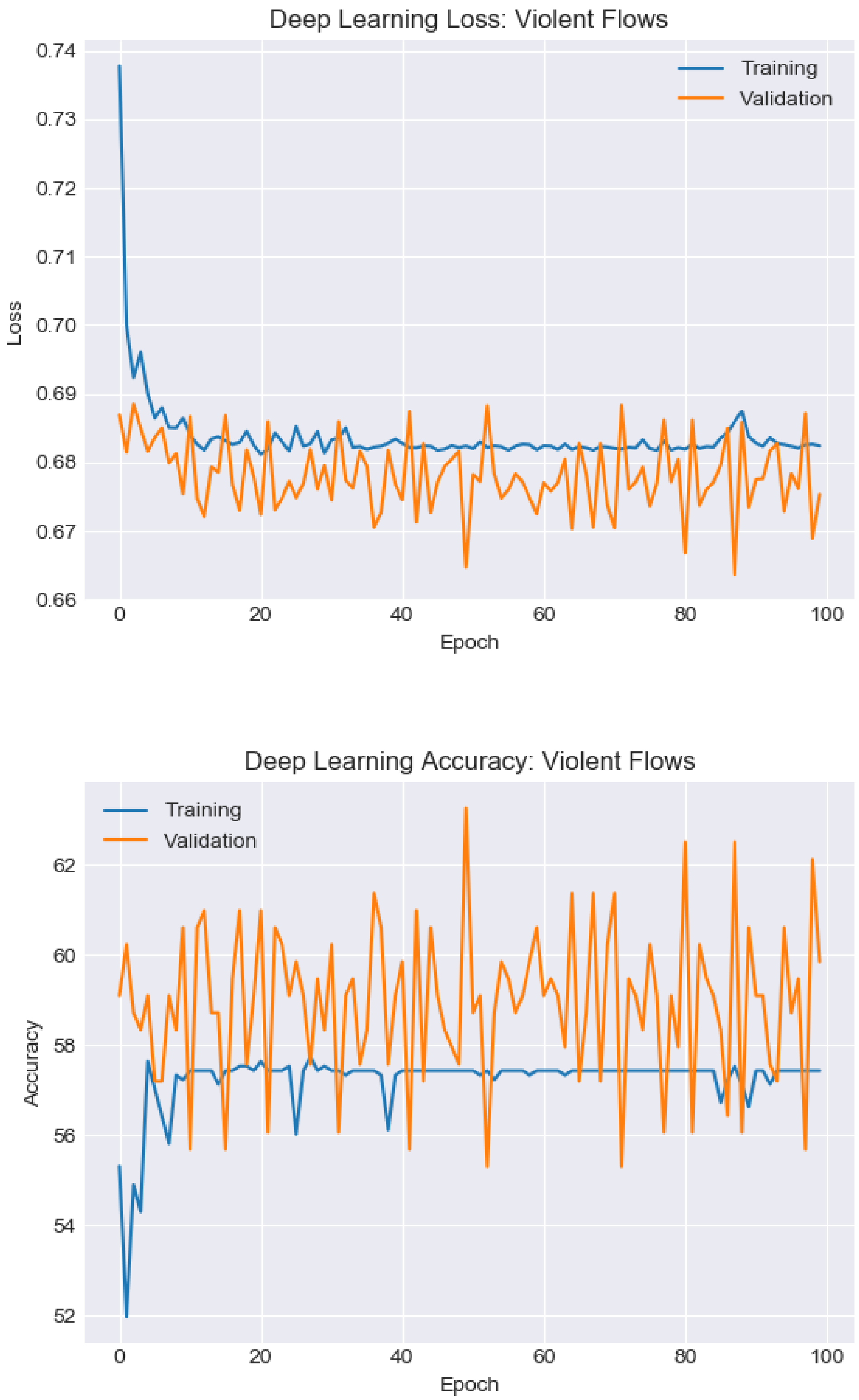


Figure 5.14: Training and validation accuracy/loss when fitting a C3D model using the Violent Flows dataset.

Chapter 6

Predicting Violent Hotspots

6.1 Introduction

Across Britain, there exists a night-time social culture in which groups of people congregate around establishments that serve alcoholic beverages. There is an increased risk of assault-related injury within and around premises licensed to sell alcohol [149]. Increased levels of drunkenness, disorderly behaviour and assault-related injury, characterise the environment associated with this social culture [44, 82, 83, 161]. Information regarding the practices observed in a social setting can be used to inform policing strategies, for instance, in response to the observed increase of injury around licensed premises, targeted policing procedures were applied, resulting in a substantial reduction in assaults [149]. Developing policing strategies using information about the local area and its communities leads to a reduction in the severity of wounds inflicted during a violent incident [37]. Minimizing the effects of violence has a substantial impact on the related costs, such as medical costs, and costs associated with lost working hours [38]. Numerous factors are assumed to influence harm in the Night-Time Economy (NTE), including premises opening hours, congestion, population, blood alcohol level, etc.

An ABM was developed by Moore *et al.* [101] that saw the implementation of a drunken stagger to simulate intoxicated behaviour associated with pedestrians that have ingested alcohol. The gait of a pedestrian varies considerably depending on their level of alcohol-induced intoxication [1, 110]. Alcohol affected gait causes greater uncertainty in pedestrian flow around bars and nightclubs [101]. In this chapter, simulation modelling is used to investigate the effects of alcohol affected gait on the emergence of violence in the NTE. The hypotheses being that greater intoxication disrupts normal emergent behaviour seen in sober pedestrians, giving rise to encounters that, by their nature, increase the likelihood of violence and aggression. The significant value of ABMs is that they can be used in a variety of contexts to provide risk assessments and inform environment design to minimise harm. ABMs have not been rigorously applied to NTE contexts, and it is not clear whether simulations can provide insights to inform the design of the NTE. The aim of the work presented in this chapter is to investigate the relationship between observed pedestrian behaviour produced by an ABM and real-world crime data. The purpose of this analysis is to evaluate the suitability of an ABM as a violent behaviour prediction tool. Additionally, with respect to real-world environments, a robust comparison between Moore's intoxication informed ABM, and a sober model is performed.

Understanding where and when disorder and violence will occur is of value to efforts aimed at mitigating harm: this can be achieved by altering physical or social environments through means such as pedestrianising streets, or by adjusting the opening times of pubs, clubs, and bars. Agent-based models provide in silico simulations of real-world pedestrian flow but assume agents are homogeneous. In many night-time locations, alcohol is used as a social lubricant and is synonymous with an unsteady gait and violence. Empirically justified effects of alcohol, by time and dose, are incorporated in an agent-based model to simulate drunken behaviour. Geo-coded crime data is used to evaluate and compare violent crime modelling using intoxicated and sober ABM simulations. In addition to intoxication state, the relationship between invasions of personal space, pedestrian density, and pedestrian velocity with violent behaviour are investigated.

The work in this chapter demonstrates that an agent-based model can be used to predict violent crime, and that simulating intoxicated pedestrian dynamics can be beneficial for violent crime modelling. Additionally, the development and inclusion of a measure of PSI was informed by the literature, and determined to be a useful factor for the prediction of violent crime.

6.2 Intoxicated Agent-Based Modelling

Presented in this section are the implementation details for the ABM used to investigate the correlation between emergent behaviour and real-world crime data. The proposed investigation requires a model that simulates pedestrian movement to, and from, drinking establishments in an NTE environment. Implementation details are presented across two sub-sections, the first discusses inter-agent interactions, and the second describes route selection. The implementation details introduced combine to produce a model in which an agent moves through an environment in a manner that is representative of either a intoxicated or sober pedestrian.

6.2.1 Intoxicated Gait Simulation

Work presented by Moore *et al.* [101] defined intoxication (drunkenness) as a loss of balance leading to the inability of the intoxicated person to keep their centre of mass stable. This behaviour is simulated by applying a propulsion force generated by randomly sampling a standard Gaussian distribution ($\mu = 0$, $\sigma = 1$). The simulated force associated with intoxication F^{drunk} is applied to a pedestrian at each time-step t as expressed in Equation 6.1. To avoid sharp jittering motion caused by large random forces, the force at time-step t, F_t^{drunk}, is smoothed using

F_{t-1}^{drunk}.

$$F_t^{\text{drunk}} = 0.5F_{\text{rnd}} + 0.5F_{t-1}^{\text{drunk}} \tag{6.1}$$

The Generalized Centrifugal Force Model (GCFM) is a spatially continuous force based model for pedestrian simulation. As typical with force based models, the movement vector of a pedestrian is determined by the sum of forces (Equation 6.2); these being the driving force $\overrightarrow{F_i^{\text{drv}}}$ for agent i, repulsion force $\overrightarrow{F_{ij}^{\text{rep}}}$ between agent i and neighbouring agents j, and repulsion force $\overrightarrow{F_{iw}^{\text{rep}}}$ between agent i and wall w. For intoxicated pedestrian simulation, the GCFM force formulation is modified by adding an intoxicated force (Equation 6.3).

$$\overrightarrow{F} = \overrightarrow{F_i^{\text{drv}}} + \sum_{j \in N_i} \overrightarrow{F_{ij}^{\text{rep}}} + \sum_{jw \in W_i} \overrightarrow{F_{iw}^{\text{rep}}} \tag{6.2}$$

$$\overrightarrow{F} = \overrightarrow{F_i^{\text{drv}}} + \sum_{j \in N_i} \overrightarrow{F_{ij}^{\text{rep}}} + \sum_{jw \in W_i} \overrightarrow{F_{iw}^{\text{rep}}} + \overrightarrow{F^{\text{drunk}}} \tag{6.3}$$

6.2.2 Agent Routing

To guide an agent from one point in space to another, a traversal path is selected. A traversal path describes a set of points between an agents' current position and its goal. Points along a traversal path are used to guide an agent towards its destination. During the initialisation stage of our simulation process, a graph is generated. Each edge in the graph denotes the cost of traversing from one point to another, where the cost of traversal is defined as being equal in value to distance in metres. Using a graph-based search, the optimal traversal path between two points in an environment can be identified. To generate behaviour that is more representative of real-world environments, a road crossing deterrence is implemented to keep agents on the pavement, and a random process is added to increase variation in the traversal path selection process.

Road Walking Deterrence

Our experiments involve simulating NTE environments, these environments include roads with no marked or signalled crossing points, allowing pedestrians the ability to cross the road whenever they deem appropriate. Palamarthy *et al.* [108] used Gap Acceptance theory to decide when to cross a road, while Wang *et al.* [147] studied the factors associated with the unmarked crossing of roads. Unfortunately, research that informs *where*, rather than *when* to cross does not exist at the time of writing. Data describing traffic conditions in our simulated environments in unavailable. Given the lack of required information, vehicle movement or traffic is not explicitly simulated; instead, it is assumed that pedestrians in a simulation expect vehicles to be on the road, and as a result, aim to remain on the pavement to avoid accidents. The lack of functional vehicle simulation does not allow for emergent behaviours such as pedestrians waiting on the pavement until it is safe to cross.

To ensure pedestrians act realistically and stay on the pavement, a road crossing penalty is applied to an agents' path choice when they select their route. A road crossing penalty is induced based on the amount of time a pedestrian would remain on the road. Typically, the cost associated with a pedestrian moving from one point to another is equal to the distance of the journey in metres. However, when crossing a road, an increased traversal cost is induced. The road traversal cost is equal to the product of the distance d of crossing and the angle of incidence θ (in radians). The cost of traversing a road is defined as $d(\theta + 1)$; the addition of 1 ensures that the cost associated with crossing a road equals at least d, the standard traversal cost between two points. Figure 6.1 demonstrates the angle of incidence. The road traversal cost is lower when the path of motion of a pedestrian is closer to being perpendicular to the direction of the road.

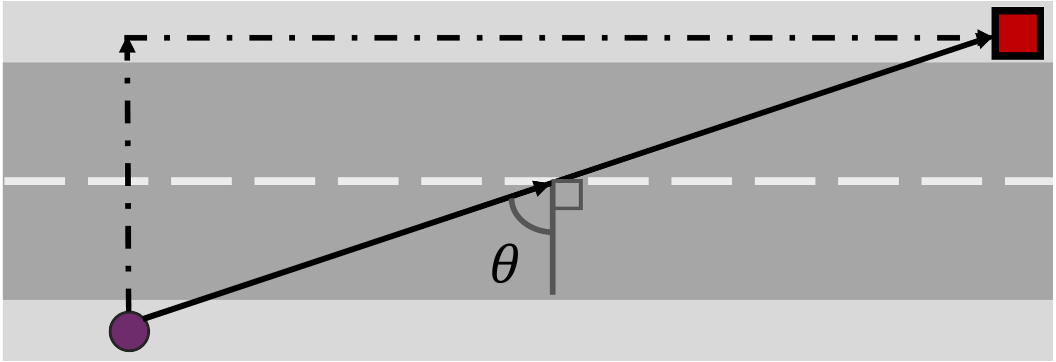


Figure 6.1: Theta is the minimum angle between the pedestrian's direction of travel and a line perpendicular to the direction of the street. The red square is the intended destination for the purple pedestrian.

Random Path Selection

If all pedestrians in our simulation selected the optimal traversal path, the we would observe uncharacteristic behaviour. It would be unrealistic to expect every person, especially while intoxicated, to make optimal path selection decisions. Given the lack of literature on the relationship between intoxication and routing decisions, the following decision was made arbitrarily. When routing, a pedestrian selects the top 10% of the possible traversal paths leading from a given point to the pedestrians goal, and then selects one randomly.

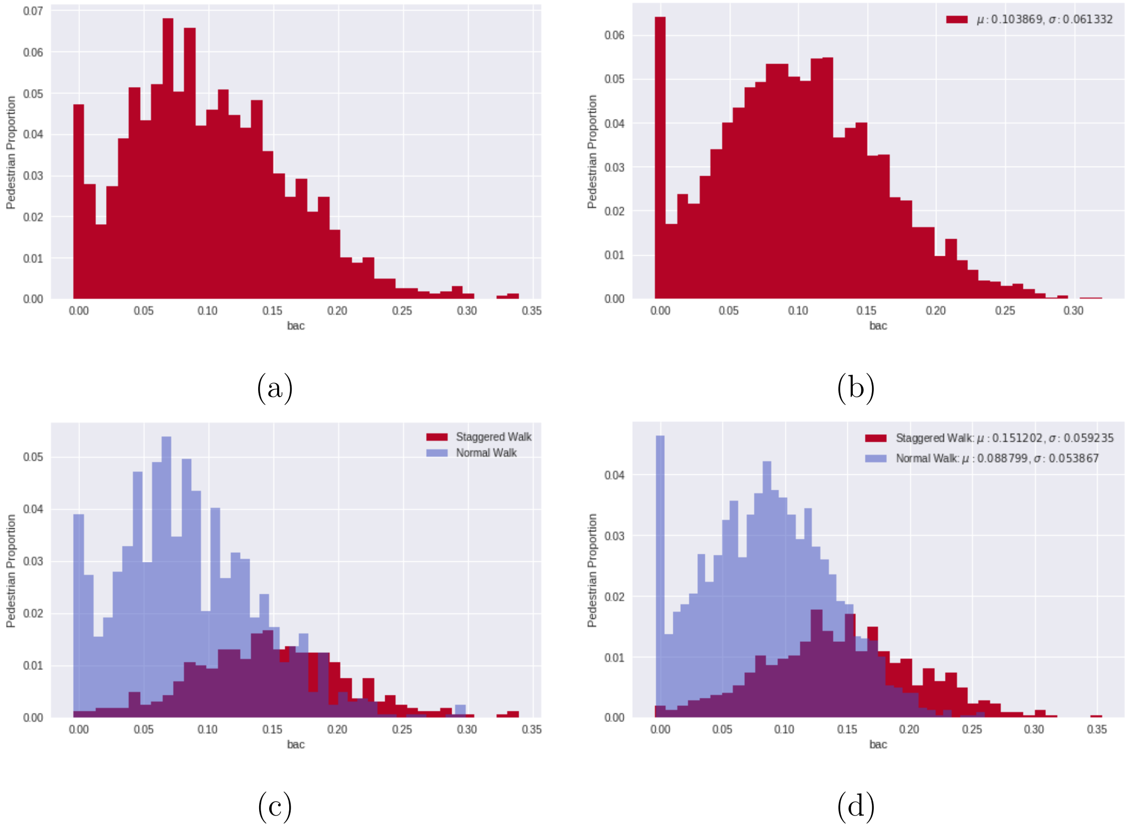
6.3 Data

GPS crime data and CCTV surveillance data was obtained from both Northampton and South Wales Police forces. The data included with the GPS reports were *time*, *longitude*, *latitude* and *crime type*. Only crimes whose definition describes physical violence between two or more people are used. The accuracy of the GPS coordinates from Northampton is quoted to have up to a 1.5-metre deviation from the true location. No reported GPS error was provided for the South Wales Police data. Two-dimensional KDE is applied to the positional GPS crime data to generate an aerial profile of an environment; high-intensity regions represent dense crime populations; these regions highlight where violent behaviour occurs within an environment. This representation is used to evaluate how well an agent-based model can approximate true crime by comparing model output with the kernel density estimated representation of GPS crime data.

In addition to positional crime data, CCTV footage of various locations throughout the cities of Cardiff and Northampton was obtained. CCTV data is used to count the mean and standard deviation of the number of pedestrians that are participating in the NTE at any given moment in time; this value is used as an input to our simulations to ensure they reflect real-world conditions by simulating the correct number of pedestrians. Blood alcohol concentration (BAC) levels of pedestrians in the Northampton NTE were not available. Therefore, data gathered from the city of Cardiff by Perham *et al.* [110] was used to inform model drunkenness for all intoxication informed simulations. Figure 6.2(a) and 6.2(b) represent the true distribution and Gaussian approximation function used to generate the level of intoxication in drunk pedestrians for our simulations respectively.

6.3.1 Model Output

The experiments discussed in this chapter involve simulating pedestrian movement in real-world environments. Aspects associated with pedestrian behaviour are measured and correlated with real-world crime data. Three attributes associated with pedestrian behaviour are analysed, these are: density, velocity and Personal Space Invasion (PSI). As stated in the Section 6.1, existing research has shown stress to be related to crowding and violence, and that invasion of per-



(a)

(b)

(c)

(d)

Figure 6.2: True distribution of BAC levels (a) and Gaussian estimated distribution (b). Disentangled BAC distributions by stagger status, observed (c) and estimated (d)..

sonal space invokes stress in an individual. It is hypothesised that implementing a notion of personal space invasion and generating two-dimensional heat maps of PSI will produce an output that reflects true violence. The profile generation procedure presented in the JuPedSim [18] framework is used to produce a two-dimensional profile that represents an environment from a top-down perspective, where intense invasions of personal space are represented using greater values. To generate this data, each agent in a simulation is assigned a value P that indicates the extent to which an agents' personal space is invaded. P (Equation 6.4) is expressed as the sum of exponential distances between an agent and all its neighbours within a 1.2 metre radius, the boundary of personal space defined by Hall [52]. As per the JuPedSim method, a Voronoi diagram is produced for each time-step in a simulation using the positional data of each pedestrian. Each cell is weighted by the corresponding pedestrians P value. A two-dimensional profile is generated by applying a discretisation process to the weighted Voronoi diagram, where each pixel in the discrete space represents m metres in the real world. A set of profiles are compiled across many simulations and averaged to produce a single output. Examples of such outputs can be viewed in Figures 6.4 and 6.3. In addition to PSI, this process is performed to also produce an aerial profile for both pedestrian velocity and density. Simulations are performed such that the simulated time is equal to an hour of real time; this process is performed multiple times using various initialisation to remove bias induced by certain starting conditions. Generating reports is a time-consuming task, and the computation time is inversely proportional to the value of m. In Figures 6.4 and 6.3, each discrete cell represents 2m and 5m respectively; these values were chosen as they provided simulation reports with acceptable detail without requiring a large amount of time to produce.

$$\mathrm{P} = \sum_{i}^{N} exp(D_i) \tag{6.4}$$

6.3.2 Kernel Density Estimation

In their raw form, the positional nature of *violent crime* data points are incompatible with our simulation output. GPS data points exist within a continuous spatial space whereas our simulation output is discrete. Our simulations produce multiple two-dimensional arrays of values that represent a top-down view of a real-world environment, where each element in the array represents an area in the real-world in m^2 units. To perform correlation and prediction of violent crime, we need to place our crime data points into the same spatial structure.

To generate our ground truth data from GPS data points, Kernel Density Es-

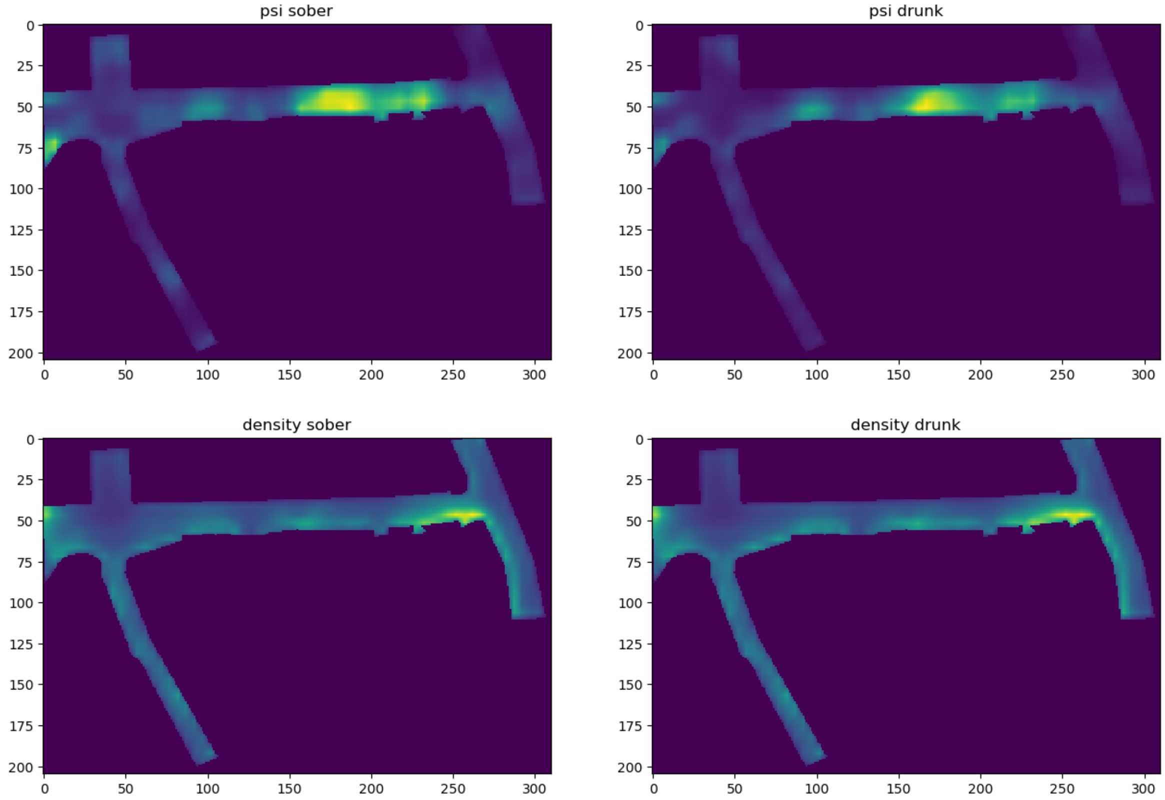


Figure 6.3: Heatmaps produced post simulation. Results are averaged across 50 initialisations.

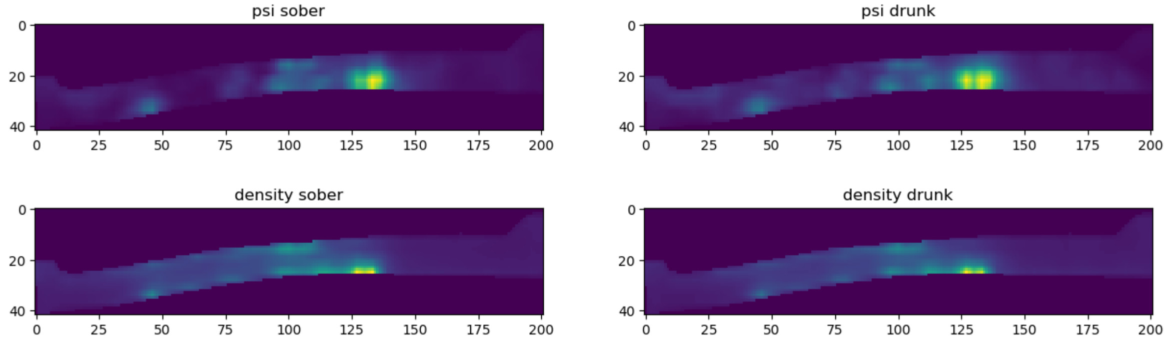


Figure 6.4: Heatmaps produced post simulation. Results are averaged across 50 initialisations.

timation (KDE) [126] is used. The KDE process will produce a two-dimensional probability density function of violent crime that is sampled using a discrete grid with dimensions that match the simulation output. There are two parameters that define the nature of the resulting probability density function, these are the kernel type and bandwidth. It is assumed that the distribution of violent crime is Gaussian in nature, and a Gaussian kernel is used when applying KDE. The bandwidth typically controls the size of the smoothing function, increasing the bandwidth will increase the size of violent crime hotspots in our ground truth data.

In the literature, there are few examples where bandwidth selection was required for analysing violent crime data [23, 141, 48]. In cases where violent crime has been investigated using KDE, a different bandwidth is selected with justification relative to the purposes of the experiment and the environment analysed. In the wider literature, Chainey [15] proposed Predictive Accuracy Index (PAI) for determining the optimal bandwidth for crime hotspot analysis. PAI is computed using two concepts, the hit rate percentage and the area percentage. The hit rate describes the proportion of crimes that were correctly predicted where an accurate prediction is defined as a crime existing spatially inside a predicted hotspot area. The area percentage represents the proportion of area that the hotspots cover with respect to the total area of the region being analysed. PAI is computed by dividing the hit rate percentage by area percentage. PAI describes the proportion of crimes predicted relative to the size of the area covered by the hotspot estimates. Bandwidth selection using PAI attempts to identify the bandwidth that produces hotspots that when analysed, results in the greatest hit rate while describing the least possible space.

To calculate PAI, 10-fold cross-validation is used to separate data into training and testing groups. Hotspots are obtained by applying Gaussian Kernel Density Estimation on the training set using a given bandwidth size. As the predicted hotspots are Gaussian, a violent crime is considered correctly predicted if it falls within the area that describes the 95% confidence interval of the estimated Gaussian. Using the test data, the hit rate percentage is computed by calculating the proportion of test data points correctly predicted using the estimated hotspots derived from the training data. The PAI computation process is performed for many different bandwidth sizes.

After applying the PAI process, the bandwidth with the highest PAI provides an associated hit rate of 23% and 66% for the Northampton and Cardiff environment respectively (Figures 6.6 & 6.5). The ground truth, in this case, would consist of hotspots that represented only 23% and 66% of the crime data within each dataset. In the case of Northampton, the majority of violent crime data samples will be excluded from the analysis. To produce an accurate predictive model, our ground truth must represent as much data as possible. A constraint is applied to ensure a bandwidth is selected such that predictive ability (hit rate) is above 90%. The hit rate constraint can be adjusted based on the requirements of the user, if a less accurate model is sufficient, then a lower value can be selected. A high hit-rate threshold may result in a bandwidth whose associated hotspots are enlarged such that they consume the entire environment; in this case, a perfect hit rate would be achieved, but all information regarding the spatial distribution would be lost. A value of 90% was selected so that the ground truth is representative of a substantial majority of the data available. The bandwidth associated with the highest PAI value with a hit rate greater than 90% is 0.01508 for Northampton and 0.05578 for Cardiff. Figures 6.7 and 6.8 depict the estimated violent crime hotspots for their respective environment using the *constrained PAI* method of bandwidth selection.

6.4 Experiments

In this section, two distinct NTE environments from two cities in the United Kingdom are described. Each environment is simulated using sober pedestrian dynamics and intoxicated pedestrian dynamics. Each model, sober and intoxic-

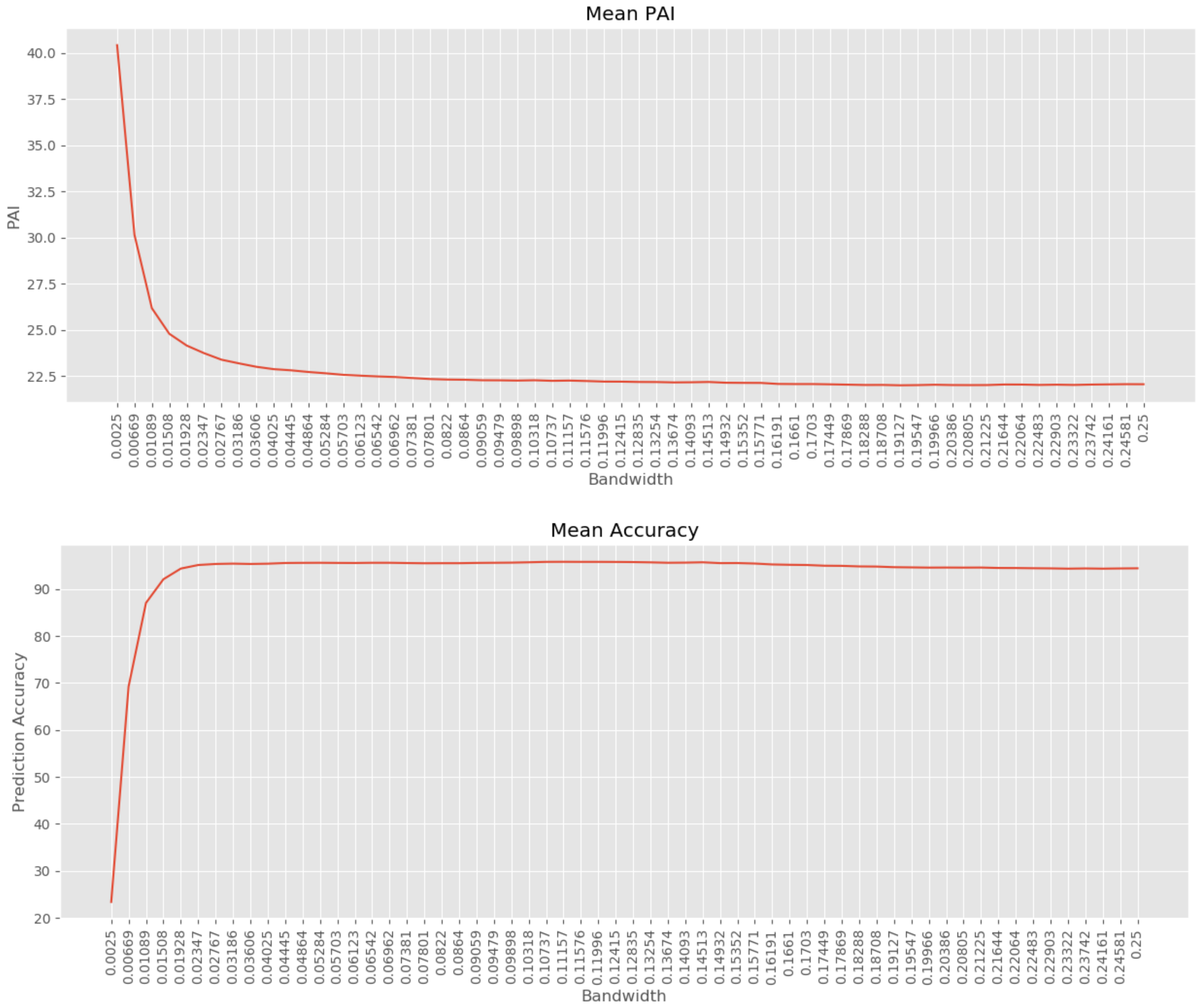


Figure 6.5: PAI and predictive accuracy (hit rate) for each bandwidth applied to the Northampton data.

ated, is evaluated with respect to their ability to generate an output that can predict violent crime.

6.4.1 Case 1: Northampton Club District

The first experiment outlined in this chapter focuses on Northampton's clubbing district. This particular environment is constrained to a single street (Bridge

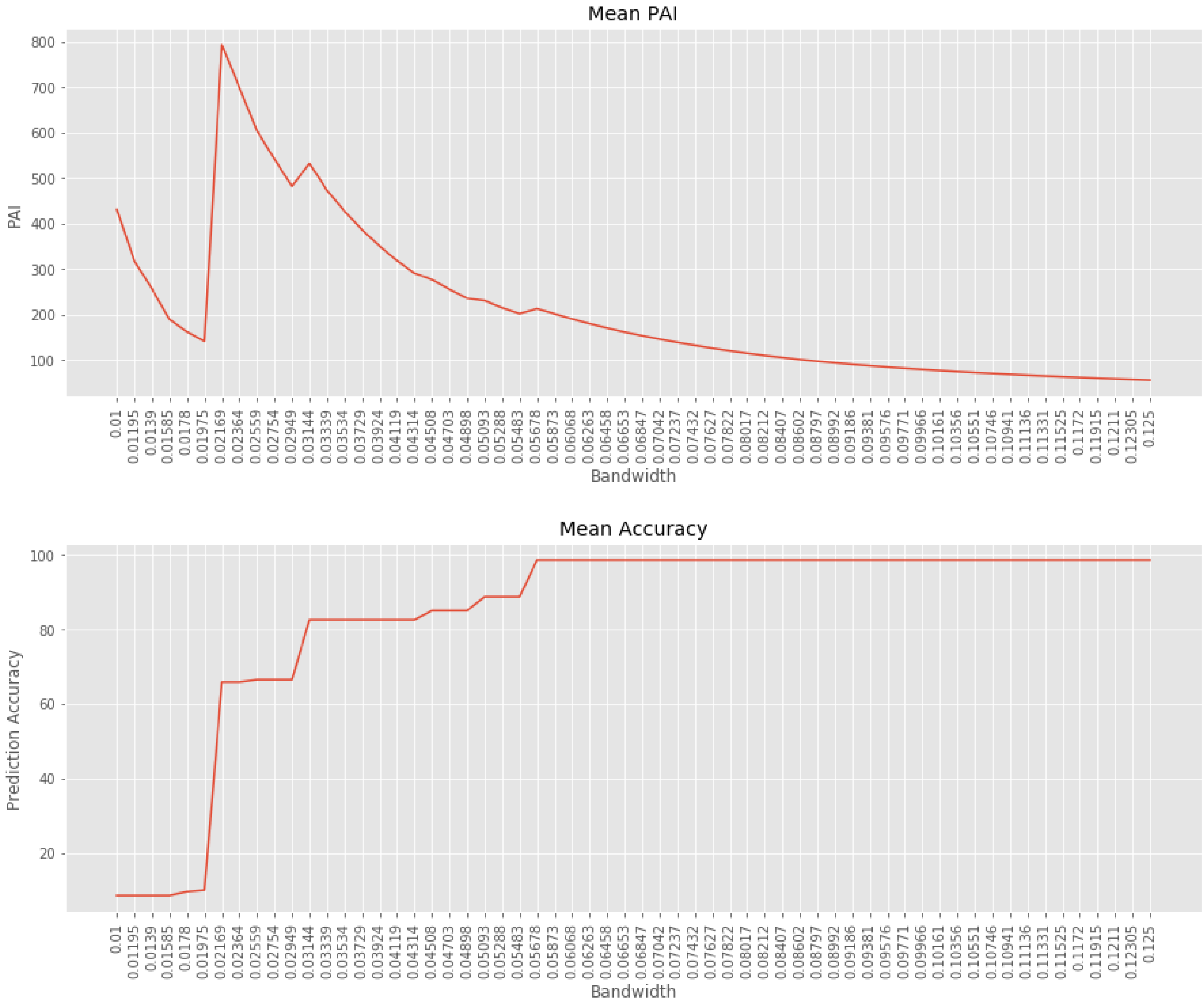
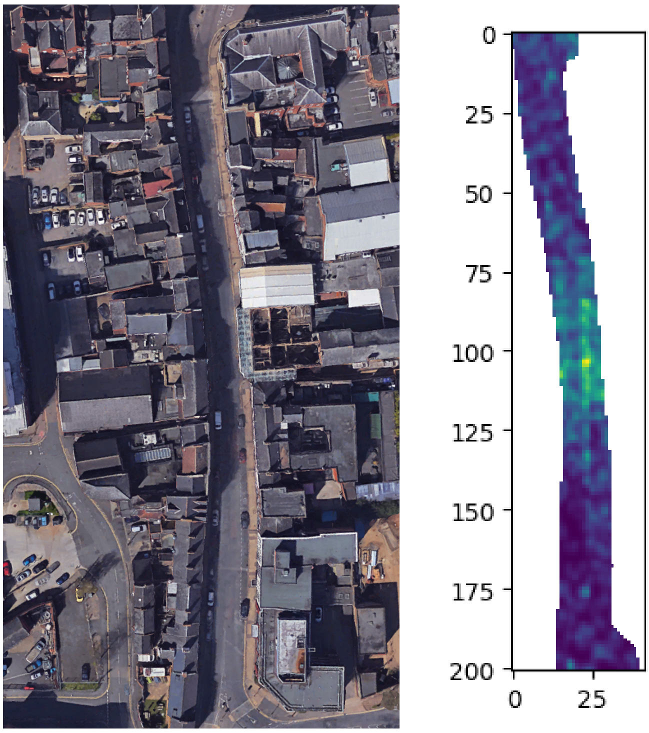


Figure 6.6: PAI and predictive accuracy (hit rate) for each bandwidth applied to the Cardiff data.

Street) as all clubs within the local area lay within it. There is a total of five establishments that serve alcohol during NTE hours. The complexity of the simulated environment is low, as its design is linear as shown in Figure 6.7. Accompanied by the aerial view of the simulated environment is a profile view representing true violence (Figure 6.7). A functioning road that allows traffic to flow unconstrained at any time is present. Our simulations account for this and aim to keep pedestrians walking along the pavement as observed in CCTV



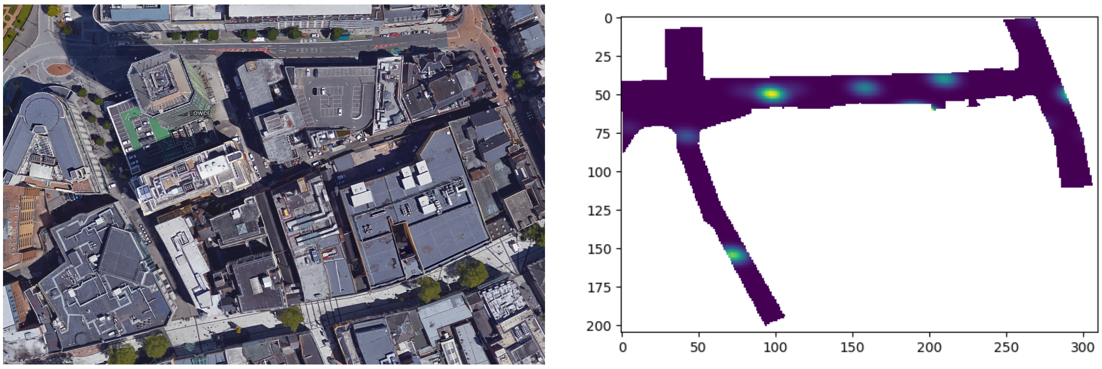
(a) (b)

Figure 6.7: a) Aerial view of the Northampton club district. b) Ground truth violent crime distribution (PAI selected KDE bandwidth).

footage of the environment. The average target population of our simulation is 64.369, with a standard deviation of 10.141, and we add agents to our simulation such that the agent count matches the true average pedestrian count.

6.4.2 Case 2: Cardiff's Greyfriars Road

The second experiment is reflects on one of the main clubbing districts in the city of Cardiff, Wales. A notable aspect of this environment as displayed in Figure 6.8 is that the environment is non-linear, requiring pedestrians to traverse corners. The primary area of simulation, in this case, is shared between two streets, Greyfriars Road and The Friary, with five and two drinking establishments respectively. Along The Friary no cars are permitted, creating an open area for pedestrians to walk freely. Along Greyfriars Road, the road accepts traffic, therefore, the road crossing behaviour discussed in Section 6.2.2 is applied. The average target population of our simulation is 100.161, with a standard deviation of 16.168, as guided by real-world pedestrian counts of the area.



(a)

(b)

Figure 6.8: a) Aerial view of the Greyfriars road. b) Ground truth violent crime distribution (PAI selected KDE bandwidth).

6.5 Results

Correlation measures and Ordinary Least Squares (OLS) regression are used to evaluate and compared pedestrian simulations. Correlation analysis using Zou's test [163] is performed to determine whether the output produced by an intoxication informed model is significantly different from a standard sober model. Regression modelling is used to generate a prediction model that utilises all three measures of pedestrian behaviour. A regression model is used to determine the importance of factors for modelling violent crime. More specifically, regression is used test the impact of PSI for modelling violent crime when considered alongside measurements of density and velocity.

6.5.1 Correlation

In this section, the correlation between individual measurements of pedestrian behaviour and violent crime is examined. The aim is to determine whether the correlations reported using measurements from a intoxication informed model are significantly different from correlations between true crime and measurements output from a sober pedestrian simulation model.

Both measures of personal space invasion (PSI) and density correlate positively with violent crime (Figures 6.9 & 6.10). A weak inverse correlation between violent crime and pedestrian velocity is also observed. The correlation strength for density is greater than PSI for the Northampton environment; the opposite is true for the Cardiff environment.

The Zou confidence intervals indicate whether two correlation values are significantly different as well as the direction and magnitude of the difference.If the confidence interval contains zero, then the observed difference is insignificant, and that the difference in correlation values cannot be declared as originating from two distinct processes. Pearson's R and Spearman's Rho correlation measures are used for the Northampton and Cardiff environments respectively; these correlation methods were selected by analysing the bivariate scatter plots of density/PSI/velocity and violent crime.

The bivariate scatter plots associated with the Northampton experiment show a linear relationship between variables of density, PSI and velocity with violent crime (Appendix D.1). The Zou confidence intervals show the correlation values for each variable is significantly different between intoxicated simulation and non-intoxicated simulation. The correlation values derived from the intoxicated process are weaker than the same correlations reported by the non-intoxicated process.

The bivariate scatter plots associated with the Cardiff simulation show a non-linear relationship between variables of density, PSI and velocity with violent crime (Appendix D.2). Zou confidence intervals reveal that the difference between PSI correlation values for intoxicated and non-intoxicated models is insignificant at lower KDE bandwidth values; when the bandwidth is greater than 0.05, the difference becomes significant and positive, indicating that in terms of PSI, the intoxicated simulation better correlates with violent crime than the non-intoxicated process. At low KDE bandwidth values (< 0.05), the difference in density correlation values is significant and negative, indicating that at low KDE values, the non-intoxicated simulation produces a density output that more strongly correlates with violent crime than the intoxicated process. At higher KDE bandwidth values (> 0.10), the difference between significant and positive. At the bandwidth determined by PAI, the difference between PSI correlation values is significant and positive, and the difference between density is significant and negative.

6.5.2 Regression

To understand the factors that are important when predicting violent crime, an OLS regression model [75] is used. The purpose of this analysis is to compare sober and intoxication informed simulation output, and to also determine whether PSI provides any meaningful information for the task of predicting violent crime when considered alongside measures of density and velocity. In this experiment, the predictor variables are density, velocity and PSI. The target variable for the OLS regression is a true crime heat map produced by applying the PAI method of bandwidth selection to GPS crime data. The determined kernel bandwidths are 0.032 and 0.045 for the Northampton and Cardiff experiments respectively.

Four models based on the environment (Cardiff/Northampton) and pedestrian intoxication state (Sober/Drunk) are produced. The Northampton environment is associated with a multicollinearity index of 16.29 and 10.04 for the sober and

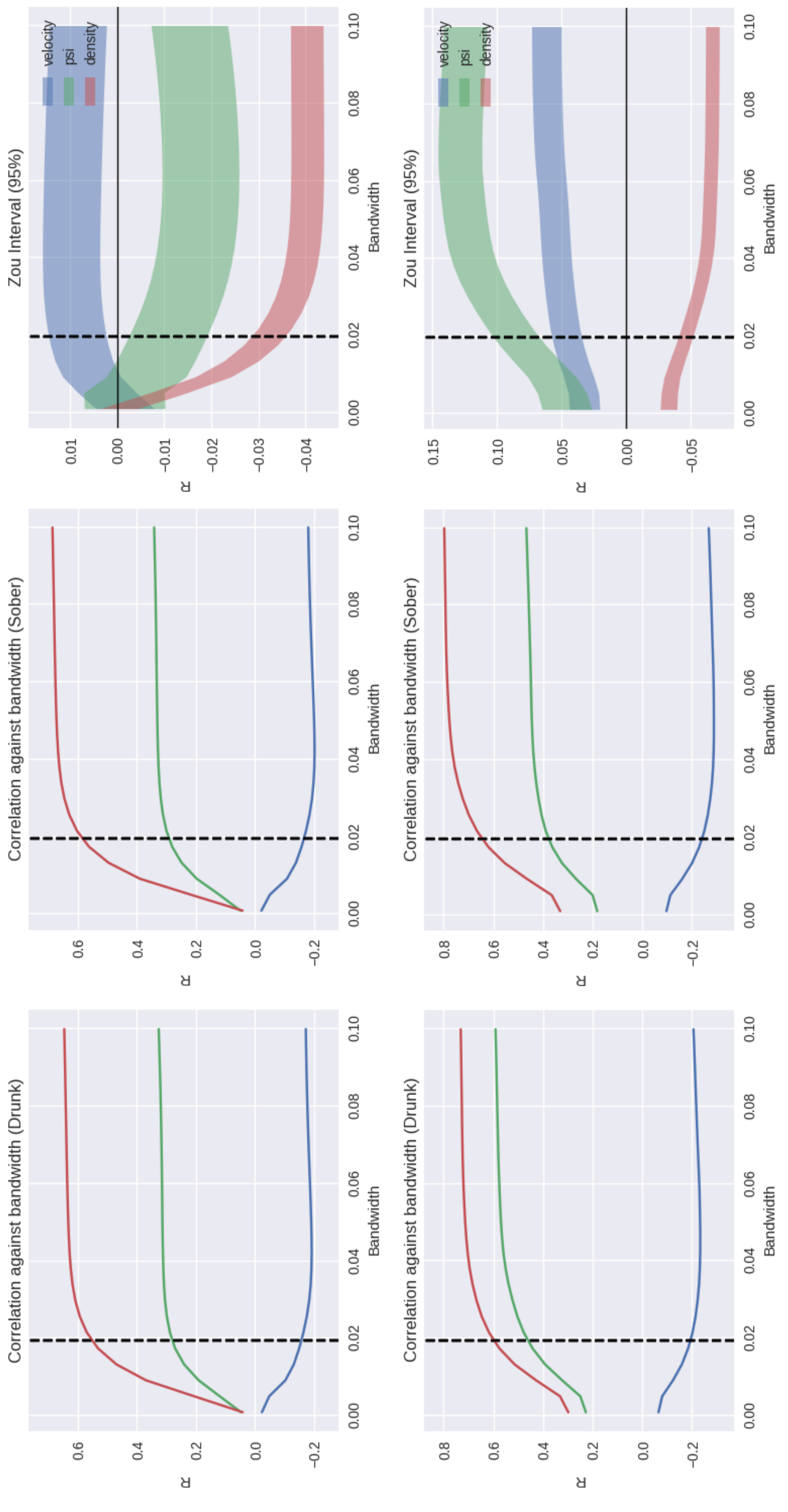


Figure 6.9: Pearson(top) and Spearman(bottom) correlation between Velocity, Density and PSI with Violent Crime. Northampton experiment.

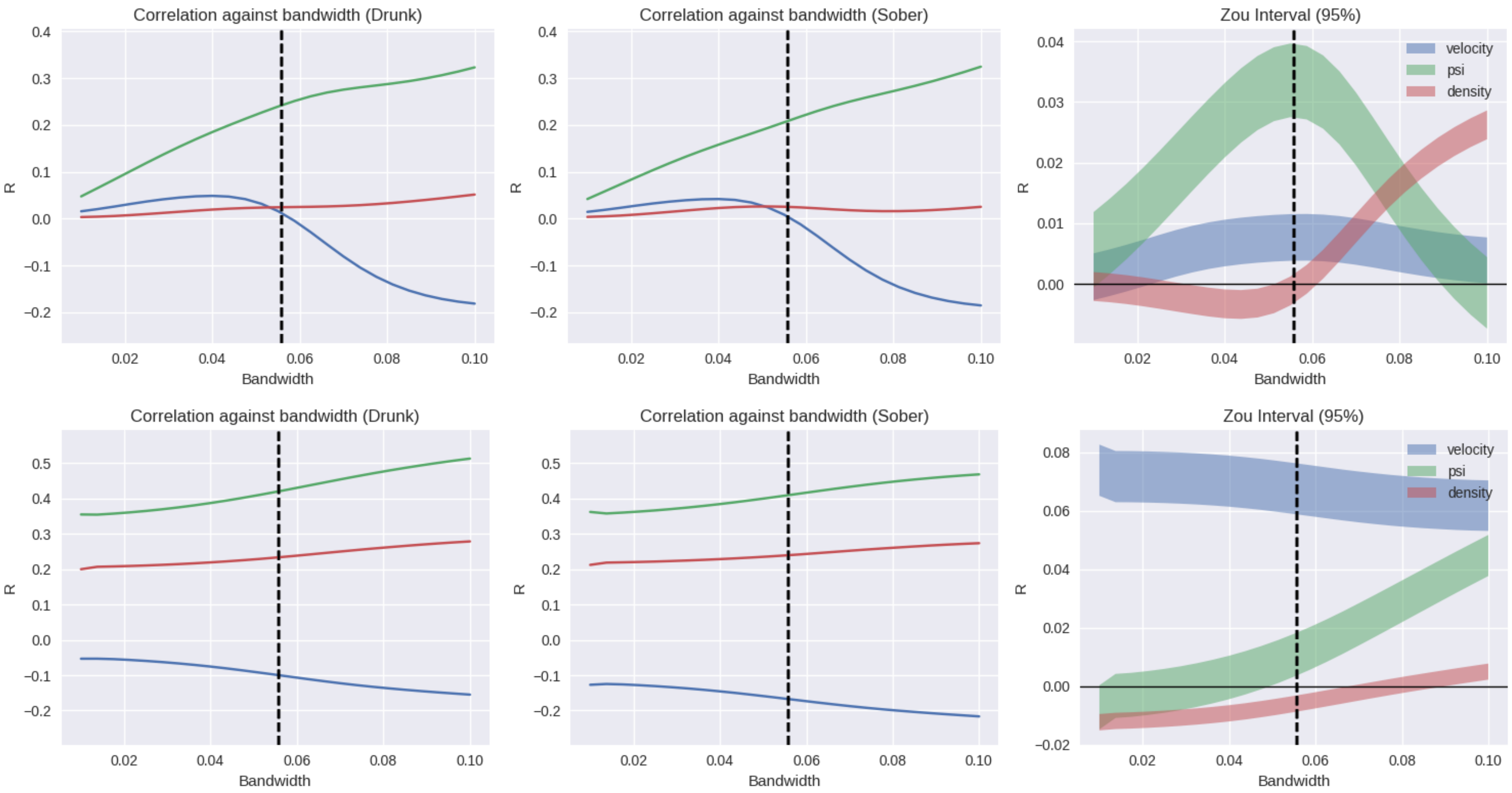


Figure 6.10: Pearson(top) and Spearman(bottom) correlation between Velocity, Density and PSI with Violent Crime. Cardiff experiment.

intoxication informed variables respectively. Regarding the Cardiff environment, the multicollinearity index values are 15.96 and 15.38 for sober and intoxication informed variables. Given that the multicollinearity index is high, the interpretation of the OLS model factors can be imprecise. To address this, Principal Component Analysis is used to transform the data such that the multi-collinearity index is 1. The transformed data is less interpretable as the information from one factor can be distributed across multiple principal components. Tables 6.1 and 6.2 present the loadings and squared correlation between the original predictor variables and the transformed principal components. The importance of predictors and their relationship with violent crime can be interpreted using PCA variable loadings and regression coefficients.

		Squared Correlations			Loadings		
Model	Data	PC1	PC2	PC3	PC1	PC2	PC3
Sober	Velocity	0.329298	0.670215	0.000486	-0.419034	0.907106	0.0396252
	PSI	0.783383	0.058214	0.158403	0.646312	0.267341	0.71471
	Density	0.762703	0.0860848	0.151212	0.637724	0.325098	-0.698298
Intoxicated	Velocity	0.342606	0.657165	0.000228	-0.421489	0.90637	0.0289855
	PSI	0.800706	0.0613116	0.137983	0.644354	0.276847	0.712856
	Density	0.785205	0.0814744	0.13332	0.638087	0.319138	-0.700711

Table 6.1: Correlation between Principal Components and variable loadings for the Northampton experiment.

Concerning the Northampton experiment, the squared correlations reveal that the first principal component correlates with all three variables, density, psi, and velocity (Table 6.1). The sign of the variable loadings indicate that as velocity decreases, measures of density and PSI increase. When considering the first principal component in relation to the regression coefficient (Table 6.3), the regression model suggests that PSI and density are positively related with violent crime

		Squared Correlations			Loadings		
Model	Data	PC1	PC2	PC3	PC1	PC2	PC3
Sober	Velocity	0.0945457	0.894343	0.0111116	0.26556	0.955581	0.127837
	PSI	0.602377	0.0758632	0.32176	0.670311	-0.278311	-0.687914
	Density	0.643729	0.00921382	0.347058	0.692937	-0.096992	0.714445
Intoxicated	Velocity	0.0213025	0.978122	0.000575	0.123982	-0.991816	-0.0304822
	PSI	0.684828	0.0043458	0.310825	0.702967	0.0661106	0.708143
	Density	0.679706	0.0118625	0.308432	0.700333	0.109225	-0.705411

Table 6.2: Correlation between Principal Components and variable loadings for the Cardiff experiment.

whereas velocity displays a negative relationship with violent crime. Density and PSI correlates with with the third principal component, however the correlation is much weaker than the correlation with the first principal component. The variable loadings for the third component indicate that as density increases, PSI decreases. Looking at the regression coefficient, the relationship between density and violent crime, and PSI and violent crime, with respect to the third principal component, is positive and negative respectively. To summarise this, the regression analysis across all components has revealed that the relationship between density and violent crime is positive. Regarding PSI, the relationship appears non-linear, with a portion of PSI information exhibiting a positive relationship with violent crime and a smaller proportion being negatively related with violent crime. Similar to PSI, evaluation of the variable loadings and regression coefficients reveals a non-linear relationship between velocity and violent crime.

In the case of Northampton, a regression model that considers all three variables reports an average (across all bandwidths) increase in R^2 of 7.1% and 10.3% for sober and intoxication modelling when compared to regression analysis that excludes PSI.

Regarding the Cardiff experiment, the squared correlation between velocity and the second principal component strong and weak for the first and third component (Table 6.2). The second principal component almost entirely represents velocity. The regression coefficient for the second component (Table 6.4) is negative for the intoxicated model, indicating a negative relationship between velocity and true crime. The second principal component for the non-intoxicated model is statistically insignificant ($p > 0.05$, Table 6.4). Density and PSI are positively correlated with the first principal component, and with respect to the regression coefficient, suggests that density and PSI are positively related to violent crime. The third principal component shows that as density increases, PSI decreases. With respect to the third principal component and the associated regression coefficient, PSI and density are exhibit a positive and negative relationship with true crime respectively. To summarise the findings across all components for the Cardiff environment, the relationship between PSI and violent crime is positive whereas the relationship between density and violent crime is non-linear. Depending on the intoxication state, velocity is either negatively related to violent crime or considered insignificant for prediction.

In the case of Cardiff, a regression model that considers all three variables reports an average (across all bandwidths) increase in R^2 of 88.1% and 89.3% for sober and intoxication modelling when compared to regression analysis that excludes PSI.

Similar to the analysis in Section 6.5.1, OLS regression analysis was performed multiple times, each time altering the target variable by changing the KDE kernel size used to generate ground truth. In most cases, the R^2 reported by a regression model fitted with data from an intoxication informed simulation is greater than the corresponding sober model (Figures 6.11 and 6.12). Neither intoxicated or non-intoxicated pedestrian simulations can produce measurements that predict violent crime with complete accuracy. Each model produces different outputs

Experiment	Variable	Coefficient	std.error	t-statistic	probability
Intoxicated	CONSTANT	0.00787828	7.24842e-005	108.69	0.00000
	PC1	0.00184862	5.22036e-005	35.4117	0.00000
	PC2	0.000741737	8.10551e-005	9.15102	0.00000
	PC3	-0.0042652	0.000139124	-30.6575	0.00000
Sober	CONSTANT	0.00787828	7.00965e-005	112.392	0.00000
	PC1	0.00200761	5.11942e-005	39.2156	0.00000
	PC2	0.000781503	7.76813e-005	10.0604	0.00000
	PC3	-0.00400725	0.000125896	-31.8297	0.00000

Table 6.3: Northampton regression results.

Experiment	Variable	Coefficient	std.error	t-statistic	probability
Intoxicated	CONSTANT	0.000134781	4.4355e-006	30.3868	0.00000
	PC1	0.00222796	3.76796e-006	15.1116	0.00000
	PC2	-0.00569789	4.44833e-006	-3.20604	0.00135
	PC3	0.00142708	5.6341e-006	17.5695	0.00000
Sober	CONSTANT	0.000134781	4.46858e-006	30.1619	0.00000
	PC1	5.16402e-005	3.8595e-006	13.38	0.00000
	PC2	-6.84245e-006	4.51549e-006	1.51533	0.12970
	PC3	-7.53107e-005	5.41947e-006	13.8963	0.00000

Table 6.4: Cardiff regression results.

as governed by drunken gait. Given that each simulation output is the product of a different process (intoxication state), then there exists the possibility that the information contained in the output is complementary to the task of violent crime prediction. As an experiment, regression using a joint model in which both sober and intoxication informed simulation outputs are included as variables was performed. Figures 6.11 and 6.12 indicate that model accuracy can be increased

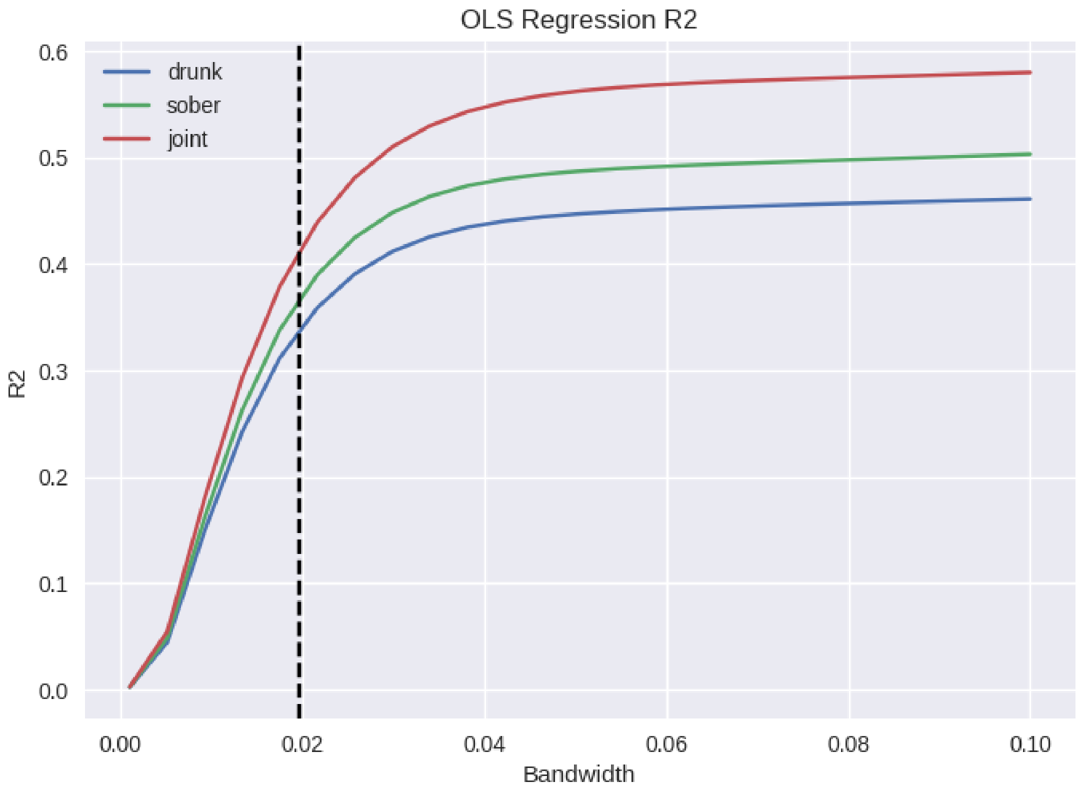


Figure 6.11: Bandwidth against OLS R squared for the Northampton environment.

by considering output from both sober and intoxication informed models at the same time.

6.6 Discussion

Based on the regression analysis, it is observed that including a measure of invasions of personal space can improve violent crime prediction. However, the degree of improvement is contingent on the environment being analysed. Our results provide evidence for the assumption that invasions of personal space induce violent behaviour. The investigation of PSI was justified by the assumption that invasion of personal space induces violent behaviour. This assumption stems

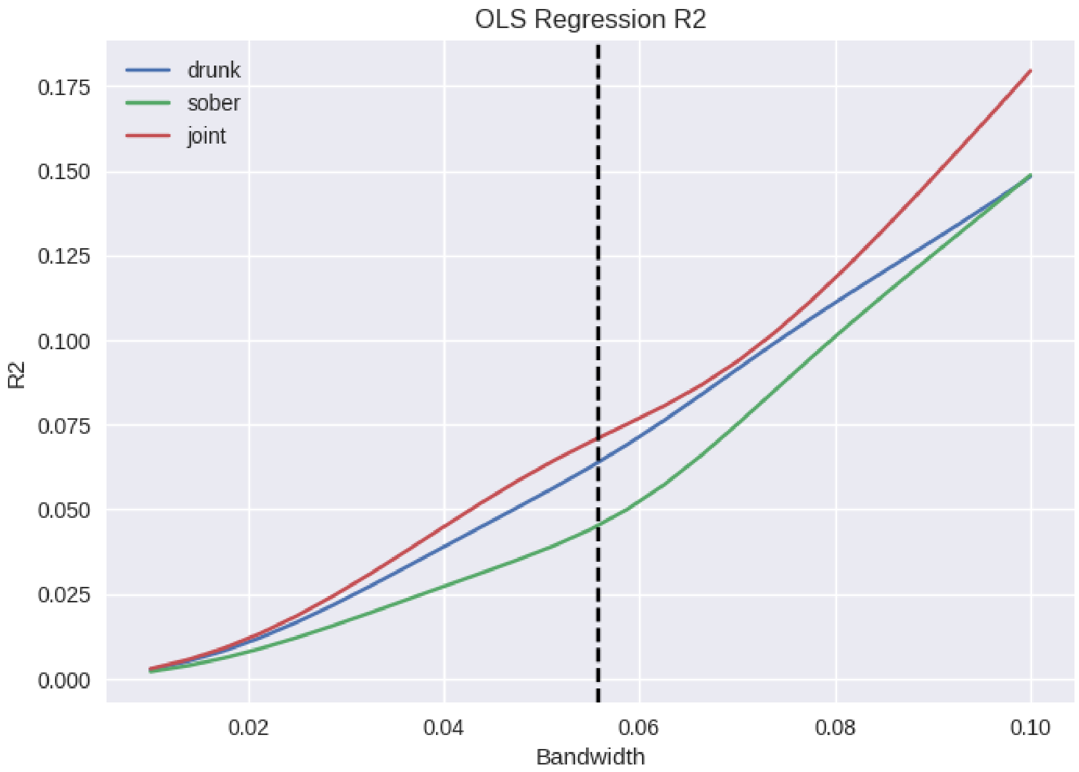


Figure 6.12: Bandwidth against OLS R squared for the Cardiff environment.

from studies by Sundstrom *et al.* [128], and Kanaga and Flynn [69] which demonstrate that invasion of personal space increases individual stress; multiple studies identify the relationship between violent behaviour and stress [4, 22, 73, 95, 103]. The experiments presented In this chapter has revealed a relationship between simulated invasions of personal space and recorded violence. Furthermore, the use of PSI as a predictor variable yields an improved model for predicting violent crime when compared to a model that excludes the variable.

Regarding the density-violence relationship, Moore *et al.* [101] state that *"it is plausible that crowding in and of itself may not be causally related to violence"*. Concerning the Northampton scenario (Section 6.4.1), it is observed that density correlates with true violence to a high degree, and achieves a greater absolute value than either velocity or PSI. Although it is plausible that crowding may not be causally related to violence, the evidence seems to suggest it can be used as a reasonably accurate indicator of violence. However, the strength of this statement does depend on the environment being analysed as indicated by the reduced performance of density when analysing the Cardiff environment (Section 6.4.2).

As of writing, there exists no literature discussing the relationship between pedestrian velocity and violence. The following results are an observation of the model's output and do not corroborate any existing research. The results show that pedestrian velocity and regions of violence have a weak inverse correlation. In the Northampton case, a strong correlation between crowding (density) with violence is observed; an associated behaviour pattern of crowding is that people move at a reduced speed as competition for space is greater. If the correlation between crowding and violence is strong then intuitively, reduced velocity may also indicate regions of potentially violent behaviour. An alternative view is that reduced velocity prolongs exposure of a pedestrian to their neighbours, increasing the likelihood that one of the neighbouring pedestrians may perform an act that induces aggression.

The output from the intoxicated model was subtracted from the output of the non-intoxicated model to highlight the difference in mode output (Figures 6.13 & 6.14). In relation to the Cardiff Environment (Figure 6.14), velocity derived from the intoxicated model is greater in regions that lack club entrances when compared to velocity derived from a non-intoxicated model. Likewise, non-intoxicated pedestrians exhibit greater velocity around club entrances than intoxicated pedestrians. This may suggest pedestrians are more easily able to resolve movement through constrained spaces when not intoxicated. In open areas, the random movement from intoxication appears to allow pedestrians to more easily resolve interactions, allowing the pedestrian to maintain a high velocity. This is in comparison to non-intoxicated pedestrians, who in the simulation, may require slowing down to avoid colliding with another person. The intoxicated model appears to increase local pedestrian density and result in increased invasions of personal space. This may result from the random intoxication force, allowing pedestrians to get closer to one another than they would normally allow when free of the influence of alcohol.

In summary, the comparison between intoxicated and non-intoxicated simulation for violent crime prediction was not conclusive. The regression analysis of the Northampton environment revealed that the intoxicated model performed worse than the non-intoxicated model, regardless of the selected bandwidth size. The correlation analysis revealed that intoxicated measures of PSI more strongly correlated with violent crime than PSI derived from a non-intoxicated simulation process. However, the opposite trend was observed for velocity and density. Concerning the Cardiff environment, the correlation analysis revealed that for small KDE bandwidths, the non-intoxicated simulation produced output that more strongly correlated with violent crime. The inverse is true when analysing violent crime heat maps generated using larger KDE values. Regression analysis reveals that the intoxicated model outperforms the non-intoxicated model across all bandwidths. For both Northampton and Cardiff environments, predictors derived from intoxicated simulation have shown to be beneficial when used in conjunction with predictors from a non-intoxicated simulation. Data from both intoxicated and non-intoxication informed simulations have proven to contribute complementary information as combining information results in a regression model with an increased R^2 score compared to a model built using exclusively intoxicated or non-intoxicated data.

Based on the presented analysis, another question arises: *Does the improvement gained from including intoxicated behaviour mechanisms in an ABM yield significant benefits in the real-world applications?* Future work is required to investigate the practical benefits of a intoxicated pedestrian model.

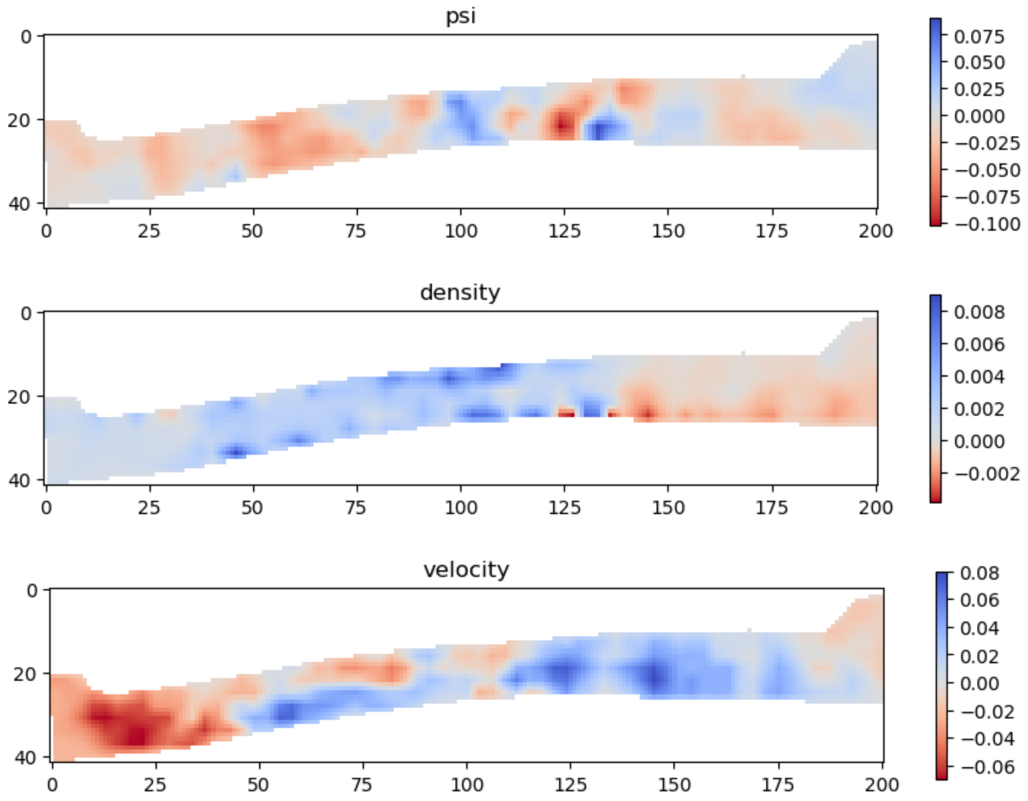


Figure 6.13: Pixel-wise difference between sober and intoxication generated output for the Northampton environment.

6.7 Conclusion

To conclude, it has been demonstrated that an agent-based model can be used to predict violent crime. Furthermore, including intoxicated pedestrian dynamics can improve predictive power when utilised alongside to a typical agent-based simulation model. Through experimentation, evidence has been provided for the hypothesis that invasions of personal space are related to violent behaviour. To achieve these findings, a measure of Personal Space Invasion was defined based on the theorem presented in the field of Proxemics [52]. Future work would involve the deployment of predictive models in the real world to inform policing strategies; by evaluating each model's ability to improve the effects of crime, it

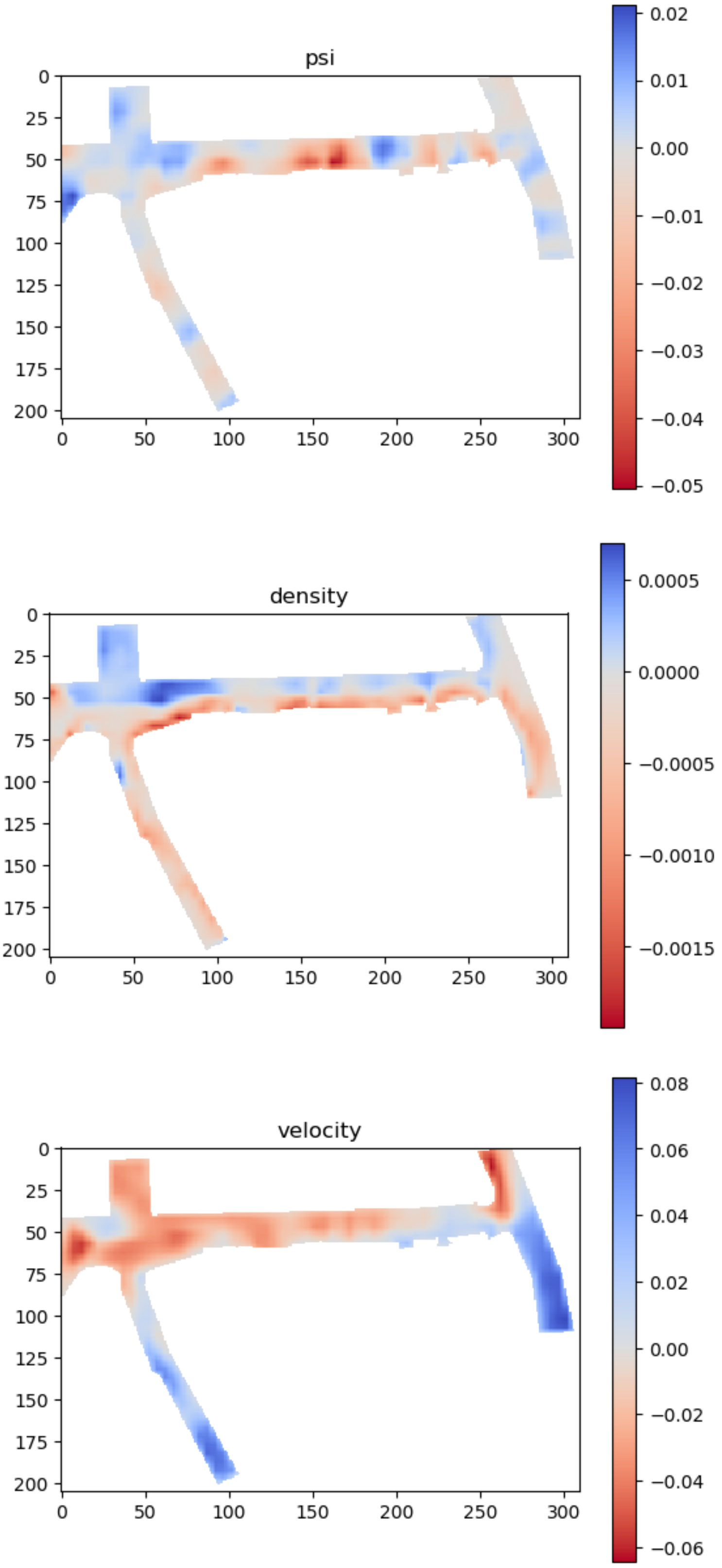


Figure 6.14: Pixel-wise difference between sober and intoxication generated output for the Cardiff environment.

will be possible to demonstrate the practical difference between models.

Chapter 7

Conclusion and Future Work

7.1 Conclusion and Future Work

Contained in this Chapter is a summary of the methods developed for this thesis along with an outline for future work. Also presented is a comparison and discussion of the two computer vision based detection methods presented in Chapters 4 and 5.

7.2 Violent Behaviour Detection using CCTV

Presented within this thesis are two methods of violent behaviour detection. The development of two methods resulted from the timing associated with the availability of data. During development of work presented in Chapter 4, data that depicted violence in crowded situations was more abundant than footage of one-on-one violence. As more data was obtained, the focus shifted towards a more general method that was theoretically more suitable for analysing one-on-one violence. Both methods demonstrated state-of-the-art classification performance at the time of their development.

Modelling Crowds using Temporal Texture

The method outlined in Chapter 4 concerns the detection of violent behaviour within crowded environments. The research disclosed in this chapter takes inspiration and methodologies from *crowd counting* techniques. Existing research has demonstrated that measurements derived from texture, specifically GLCM, are powerful at capturing structural information crowds, resulting in high crowd counting accuracy. Given that the structure of a static crowd (still image) could be encoded, it is hypothesised that temporal analysis of GLCM features could be used to capture crowd motion dynamics that are capable of discriminating between different behavioural classes when used in conjunction with a machine learning classifier. Temporal changes in appearance are represented by measuring how GLCM features change over time using four statistics: *mean*, *standard deviation*, *skewness* and *IFU*.

IFU was used to describe the stability of a crowd's appearance over time. Generally, the appearance of violent crowds between successive frames depicts greater changes in appearance when compared to normal behaviour whose appearance changes more gradually. In addition to this, it was observed that the rate of change of appearance was more variable for violent behaviour; the appearance of normal behaviour changes in a more consistent manner. Based on observation, a hypothesis was formulated and tested, that *over a short-time period, the appearance of violent behaviour is less stable over time when compared to footage of non-violent behaviour.* During the testing of this hypothesis, it was demonstrated that IFU is a powerful descriptor for use in a violent behaviour classification pipeline.

It has been argued that many methods of violent behaviour detection are too computationally expensive to be practically implemented in the real-world [49]. As discussed in Chapters 1 Chapter 2, Section 2.1, theoretical benefits to public health and safety can be achieved by using computer vision to assist in the *active* *observation* of surveillance footage. With this in mind, the algorithm presented in this chapter was designed to operate with a low computational cost.

Given that violence detection in crowds is a sub-field of abnormal behaviour detection in crowds, our method was applied to abnormality detection datasets to test generalisability. The method generalised well to other types of behaviour detection tasks, providing comparable performance with state-of-the-art methods on non-violence datasets.

Modelling One-on-One and Crowd Violence using Violent Interest Points

Presented in Chapter 5 is an alternative method that operates by locating, and subsequently describing, regions of interest based on motion characteristics associated with violent behaviour. The characteristics are: high acceleration, non-linear movement and convergent motion. A justification for each characteristic can be found within Chapter 5, but a brief overview will follow. High acceleration has long been associated with violent behaviour, a fact discussed in both Chapter 2 and Chapter 5; simply put, to incur damage, an object must be moving at high speed. The field of research associated with understanding pedestrian movement identifies that normal behaviour tends to exhibit a laminar flow, that is pedestrians typically move along a linear line without significant deviations. Finally, object convergence is an analogue of physical interaction. By computing a per-pixel measure of each of these properties, multiple response maps are generated.

Computing these characteristic on a per-pixel level results in an image where regions with greater values are more likely depict actions associated with violence. This information is used as a prior in an interest point detector scheme to produce a set of interest points based on actions that exhibit properties associated with violence. In Chapter 5, it is demonstrated that interest point sampling strategies based on violent characteristic priors produce a set of informative features. When encoded using a BoW scheme, the set of features from the interest point detector produce a more powerful description of a violent scene than features sampled using a regular grid structure.

Finally, an analysis of the underlying dynamics of violence reveals that the nature of violent behaviour can vary drastically, and that violence does not adhere to a singular definition with respect to the characteristics investigated.

Summary

A common trait and design goal of each method was the ability to operate within real-time. Real-time operation was defined as providing a decision of whether a frame depicted violence within 0.033 seconds (30 FPS); the motivation set out in Chapter 1 imposed this requirement for systems that detect violence can both decrease detection time, and increase detection rate. A system that fails to operate in real-time would fail with decreasing detection time. Additionally, detecting violence after the fact would provide no immediate benefit to the well being of those involved. However, it should be noted that analysis of archival footage may aid in court proceedings.

In summary, the main contributions associated with the detection of violence using computer vision are:

1. An analysis of the perceived bias in widely used datasets for both violent behaviour detection and abnormal behaviour detection. The experiments presented in Chapter 3 demonstrate clearly that the data capture process and capture devices have induced bias that allows for near perfect classification based on perceived characteristics of image quality and depth.
2. A computationally cheap method of violent crowd detection that operates in real-time. The proposed algorithm analyses how specific measures of texture, known to encode crowd structure, changes over time. Experiments

demonstrate state-of-the-art performance on both violent behaviour detection tasks and abnormal crowd behaviour detection tasks. Additionally, IFU was designed to measure the stability of crowd structure (appearance) over time. Using multiple datasets that contain both violent and non-violent samples, it was found that the appearance of violence is less stable over time when compared to similar scenes depicting normality.

3. Detecting points of interest using measurements associated with violent behaviour led to a method capable of state-of-the-art performance in real-time. An analysis of the three characteristics justified as being related to violence revealed that the underlying dynamics of violence change based on context. Additionally, interest point sampling outperformed dense grid sampling.

7.3 Violent Behaviour Prediction using ABM

Demonstrated in Chapter 6, an agent-based model that simulates drunken gait within real-world environments provides more accurate predictions of violent behaviour than a model that lacks drunken characteristics.

In summary, the main contributions associated with the prediction of violence using agent-based modelling are:

1. Demonstrably shown using regression and statistical testing that an agent-based model informed by drunken characteristics allows for an increased level of violent crime prediction when compared to a model that lacks drunken characteristics.

2. Regression analysis demonstrates that measurements of PSI prove useful for predicting violent behaviour, providing evidence for the hypothesis that stress-induced violence results from invasions of personal space.

Understanding where and when disorder and violence will occur is of value to efforts aimed at mitigating harm: altering the physical or social environment, such as pedestrianising streets, or changing drinking establishment opening times. Validation is required to demonstrate that changes in predicted violence that stem from modifications to the simulated environment result in similar changes in the real-world. Additionally, the current form of analysis does not investigate the spatial relationships between neighbouring regions in space. For this, spatial regression analysis is to be applied to understand whether the likelihood of violence at a given position is influenced by activity in neighbouring locations.

Included in Chapter 6 is the foundation for future work for the creation of a potentially more accurate simulation of drunken pedestrian behaviour. The data required to the realise this model is currently unavailable, and therefore the theoretical model is untested.

7.4 Merging Technologies

The predictive model of violent behaviour generates a heat map that displays, on average, locations where violent behaviour is likely to occur. However, real-world cities often host singular events that momentarily increase pedestrian density, which may influence pedestrian behaviour; the current model, based on typical behaviour may not generalise well to this case. Assuming that individual dynamics of a pedestrian movement remain constant regardless of density, then correct simulation initialisation would allow for accurate simulation of non-normal situations. Given the large quantities of cameras places around city centre environments, it is possible to determine the pedestrian quantity and even location. This information could be used to parameterise the agent-based model. Creating a system that adaptively simulates current situations would theoretically alleviate the issue of a model failing to generalise to special conditions. A real-time, adaptive heat-map of violence could then be used to feed information back to CCTV operators, who could then distribute the focus of a camera onto areas that are likely to depict violence. Cameras that are focused on areas that are less likely to capture violence could be watched by a detection algorithm. There are challenges associated with this system, the key one being pedestrian location determination. Cameras mostly capture data without depth information. Accurately reversing the project from 3D to 2D is difficult, introducing uncertainty when attempting to determine the location of a pedestrian along a street. Furthermore, occlusions also affect this process, as the camera systems may not physically observe specific pedestrians.

7.5 Future Work

In order to determine whether the methods proposed in this thesis are suitable for real-world use, more analysis must be performed. Reporting the violent behaviour detection accuracy is not sufficiently informative to suggest whether the algorithms are suitable for deployment. For instance, a study into the acceptable levels of false positives would be required. The question that is of interest is *At what point does an assistive system become detrimental to the task due to distraction caused by excessive false positives?*. The answer to this would act as an objective goal for system development; if a new algorithm produces too many false positive, then it would be considered unsuitable for real-world deployment. The information required can be obtained using a study that measures the human ability at detecting activities whilst presenting varying levels of distraction. The distraction in this case would manifest in a similar way to an assistive detection algorithm that produces false positives.

Future research could involve investigating the detection time of violence, determining how early a behaviour is identified as being violent. Violent samples in the available datasets lack the information that occurs prior to violence as well as the transition information that leads into violence. Without prior information, it is not possible to analyse the speed of detection, an important property of the motivation that guides the work presented in this thesis. Ultimately, to investigate the detection time of both humans and algorithms, more data must be gathered.

In Chapter 3, an analysis of the available data reveals that each dataset contains some sort of data bias. To further evaluate the methods proposed in Chapters 4 and 5, it would be useful to develop a testing methodology that can uncover whether feature representations exploit a data bias. A general approach to bias testing would allow for more insightful comparison between state-of-the-art approaches, as traditional measures of accuracy fail to capture issues associated with data bias.

As mentioned in Chapter 3, both the NN-Violence and CF-Violence datasets contained instances where a camera operator would focus the camera on useless information. The key example is that a camera would focus on a road, failing to capture the side-walk, the area most likely to have pedestrians. Generating a system that identifies poor camera focus could be used to direct a human operator to re-adjust in order to better maximise the capture of meaningful data. An example as to how this could be achieved would be to use a pedestrian detection algorithm. If no pedestrians are detected for an extended period, then the system would instruct the operator to readjust the camera focus. Alternatively, models of movement can be generated that encode vehicle and pedestrian movement; the direction a camera is looking can then be guided by maximising the amount of pedestrian movement in the field of view whilst also minimizing captured vehicle movement.

The work presented within this thesis does not evaluate the effectiveness of the methods at reducing the impact of violent crime, which is a key motivator outlined in Chapter 1. The methods disclosed in this thesis would theoretically reduce the effects of violence by allowing police personnel to attend a scene of interest more quickly, resulting in a demonstrable impact on injury severity. Additionally, the majority of violent instances are not observed by police, and only come to light when victims arrive at the hospital. Violence detection systems, such as those outlined in Chapters 4 and 5, should in theory aid in the identification of instances that would otherwise be missed; this does assume that an instance of violence falls within the view of a camera.

7.5.1 Drunk Pedestrian Behaviour

The model of intoxicated pedestrian behaviour used in Chapter 6 was based on an ad-hoc model of drunkenness that was visually validated. This model makes little distinction between various aspects of gait under the influence of alcohol. To inform the development of a more accurate model of drunken stagger, a survey of gait changes under the influence of alcohol was performed. However, it was found that the information required was not available throughout the few studies conducted, which gives rise to a potential avenue of research for the future. Teixido *et al.* [132] found that female stride length tends to increase and become more unstable as alcohol intake increases. The perpendicular distance between their feet as they walk decreases and become more stable. In contrast, the male stride and step-width becomes more unstable as alcohol intake increases resulting in higher variance, however the average step-width decreases. Unfortunately, this study only used two participants, and so the sample size is too small for the data to be treated as fact. Additionally, the authors note that they may not have allowed enough time for the alcohol to take full effect. Demuera *et al.* [28] performed a similar study to Teixido *et al.* [132] in which volunteers were given alcohol and instructed to walk 10 meters in a straight line. The experiment sees the ingestion of a single intake of sake (8ml/kg), which the authors estimate to produce a BAC of between 0.14 and 0.17. A key fault with this study is that BAC is estimated without regard for an individual's alcohol consumption ability, potentially providing poor correlations between true BAC and changes in gait. Jansen *et al.* [68] report that the stride length is affected significantly by the increased intake of alcohol. The authors also report that increased alcohol intake does not affect unsteadiness, or ataxia. The authors imposed a walking limit on test participants of 1.11m/s, slightly below the average walking speed of 1.4m/s. A key problem throughout each aforementioned piece of literature is that the participants do not consume enough alcohol. Perham *et al.* [110] performed breathalyser tests on people in the street during the NTE and noted some various attributes, including whether or not a person is acting violent, whether they are perceived as drunk and their BAC level. The study suggests that a BAC of approximately 0.17 or greater will result in noticeable changes in gait; this value is equal to the upper limit of participants of the 1985 Jansen study and greater than any BAC reported in the other pieces of literature. Jansen *et al.* [68] hypothesised that their participants were not drunk enough for them to observe ataxia, and given findings by Perham, this may be true. Using an approach similar to Aiello *et al.* [1] in which drunken gait is stimulated through the use of vision impairment goggles, it would be possible to gather enough data to produce more accurate data functions of alcohol consumption to changes in gait with which to better inform our simulation. Using existing research, we generated a set of data functions (Figure 7.1) that specify the change in left/right sway, and forward/backward stagger as alcohol ingestion increases. We gathered information from a range of studies and applied multiple regression techniques to obtain functions that describe characteristics of motion based on blood alcohol concentration. Unfortunately, most studies do not investigate drunk behaviour beyond 0.16 BAC, the limit after which it was reported that noticeable motion stagger is induced. More data must be gathered to complete these drunk behaviour functions.

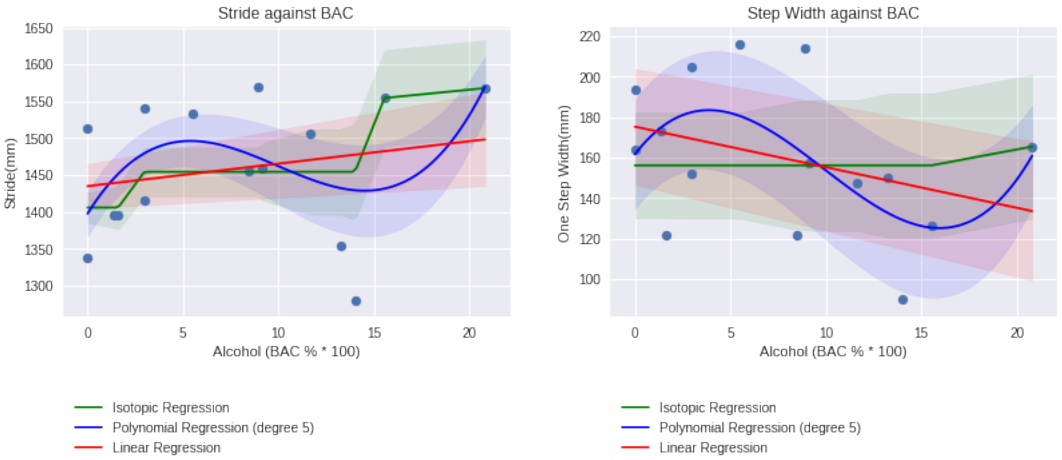


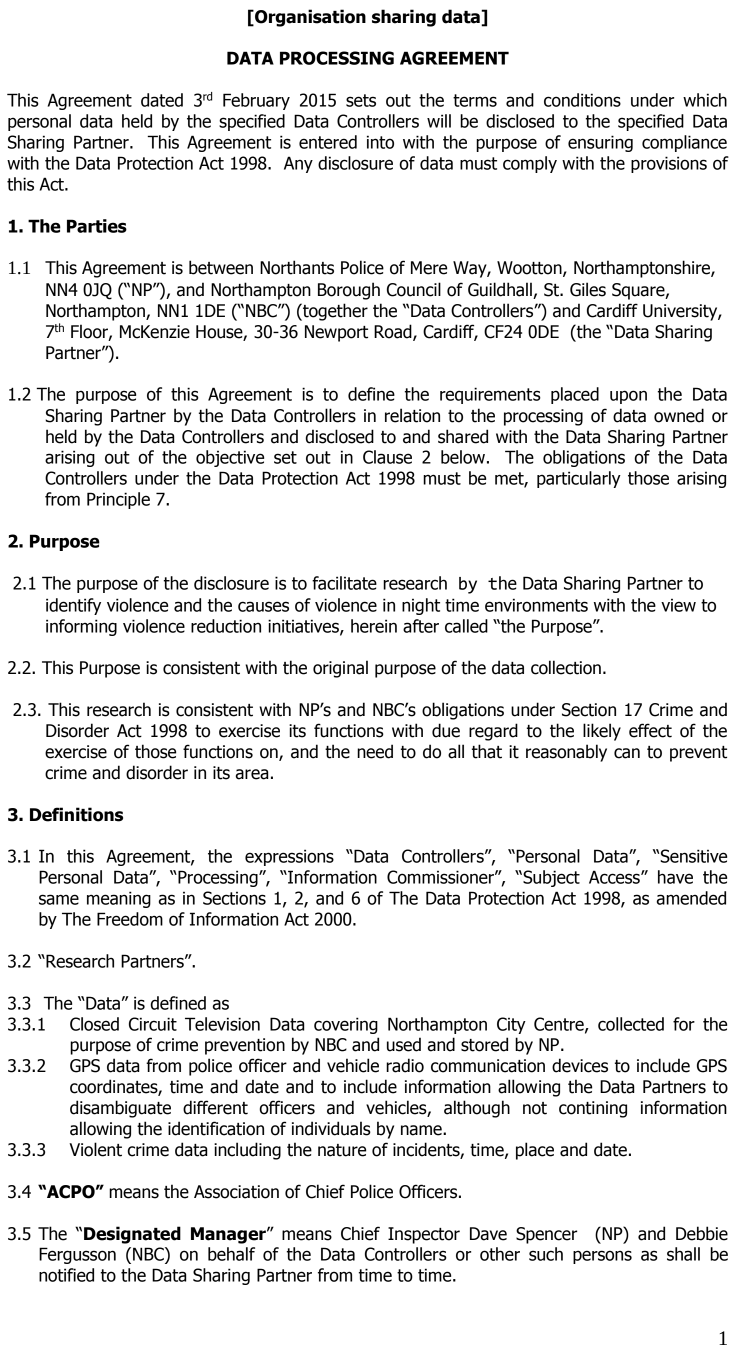
Figure 7.1: Incomplete functions visualising the change in stride length (a) and step width (b) as ingestion of alcohol increases.

Appendices

Appendix A

Data Sharing Agreement

As expressed in Chapter 3, data was obtained from Northamptonshire Police. To secure the data, a data sharing agreement was required and had to be signed by all users of the data.



[Organisation sharing data]

DATA PROCESSING AGREEMENT

This Agreement dated 3rd February 2015 sets out the terms and conditions under which personal data held by the specified Data Controllers will be disclosed to the specified Data Sharing Partner. This Agreement is entered into with the purpose of ensuring compliance with the Data Protection Act 1998. Any disclosure of data must comply with the provisions of this Act.

1. The Parties

1.1 This Agreement is between Northants Police of Mere Way, Wootton, Northamptonshire, NN4 0JQ ("NP"), and Northampton Borough Council of Guildhall, St. Giles Square, Northampton, NN1 1DE ("NBC") (together the "Data Controllers") and Cardiff University, 7th Floor, McKenzie House, 30-36 Newport Road, Cardiff, CF24 0DE (the "Data Sharing Partner").

1.2 The purpose of this Agreement is to define the requirements placed upon the Data Sharing Partner by the Data Controllers in relation to the processing of data owned or held by the Data Controllers and disclosed to and shared with the Data Sharing Partner arising out of the objective set out in Clause 2 below. The obligations of the Data Controllers under the Data Protection Act 1998 must be met, particularly those arising from Principle 7.

2. Purpose

2.1 The purpose of the disclosure is to facilitate research by the Data Sharing Partner to identify violence and the causes of violence in night time environments with the view to informing violence reduction initiatives, herein after called "the Purpose".

2.2. This Purpose is consistent with the original purpose of the data collection.

2.3. This research is consistent with NP's and NBC's obligations under Section 17 Crime and Disorder Act 1998 to exercise its functions with due regard to the likely effect of the exercise of those functions on, and the need to do all that it reasonably can to prevent crime and disorder in its area.

3. Definitions

3.1 In this Agreement, the expressions "Data Controllers", "Personal Data", "Sensitive Personal Data", "Processing", "Information Commissioner", "Subject Access" have the same meaning as in Sections 1, 2, and 6 of The Data Protection Act 1998, as amended by The Freedom of Information Act 2000.

3.2 "Research Partners".

3.3 The "Data" is defined as

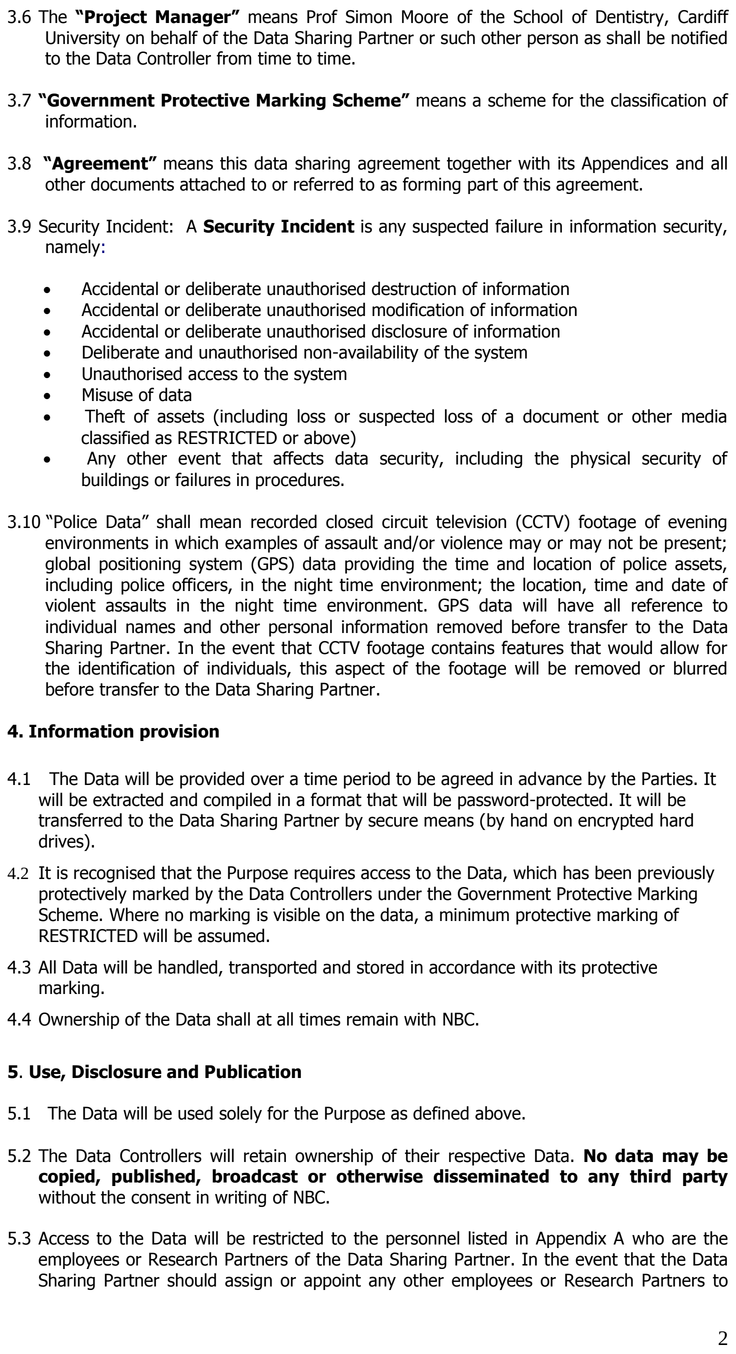
3.3.1 Closed Circuit Television Data covering Northampton City Centre, collected for the purpose of crime prevention by NBC and used and stored by NP.

3.3.2 GPS data from police officer and vehicle radio communication devices to include GPS coordinates, time and date and to include information allowing the Data Partners to disambiguate different officers and vehicles, although not contining information allowing the identification of individuals by name.

3.3.3 Violent crime data including the nature of incidents, time, place and date.

3.4 **"ACPO"** means the Association of Chief Police Officers.

3.5 The "**Designated Manager**" means Chief Inspector Dave Spencer (NP) and Debbie Fergusson (NBC) on behalf of the Data Controllers or other such persons as shall be notified to the Data Sharing Partner from time to time.



3.6 The **“Project Manager”** means Prof Simon Moore of the School of Dentistry, Cardiff University on behalf of the Data Sharing Partner or such other person as shall be notified to the Data Controller from time to time.

3.7 **“Government Protective Marking Scheme”** means a scheme for the classification of information.

3.8 **“Agreement”** means this data sharing agreement together with its Appendices and all other documents attached to or referred to as forming part of this agreement.

3.9 Security Incident: A **Security Incident** is any suspected failure in information security, namely:

- Accidental or deliberate unauthorised destruction of information
- Accidental or deliberate unauthorised modification of information
- Accidental or deliberate unauthorised disclosure of information
- Deliberate and unauthorised non-availability of the system
- Unauthorised access to the system
- Misuse of data
- Theft of assets (including loss or suspected loss of a document or other media classified as RESTRICTED or above)
- Any other event that affects data security, including the physical security of buildings or failures in procedures.

3.10 “Police Data” shall mean recorded closed circuit television (CCTV) footage of evening environments in which examples of assault and/or violence may or may not be present; global positioning system (GPS) data providing the time and location of police assets, including police officers, in the night time environment; the location, time and date of violent assaults in the night time environment. GPS data will have all reference to individual names and other personal information removed before transfer to the Data Sharing Partner. In the event that CCTV footage contains features that would allow for the identification of individuals, this aspect of the footage will be removed or blurred before transfer to the Data Sharing Partner.

4. Information provision

4.1 The Data will be provided over a time period to be agreed in advance by the Parties. It will be extracted and compiled in a format that will be password-protected. It will be transferred to the Data Sharing Partner by secure means (by hand on encrypted hard drives).

4.2 It is recognised that the Purpose requires access to the Data, which has been previously protectively marked by the Data Controllers under the Government Protective Marking Scheme. Where no marking is visible on the data, a minimum protective marking of RESTRICTED will be assumed.

4.3 All Data will be handled, transported and stored in accordance with its protective marking.

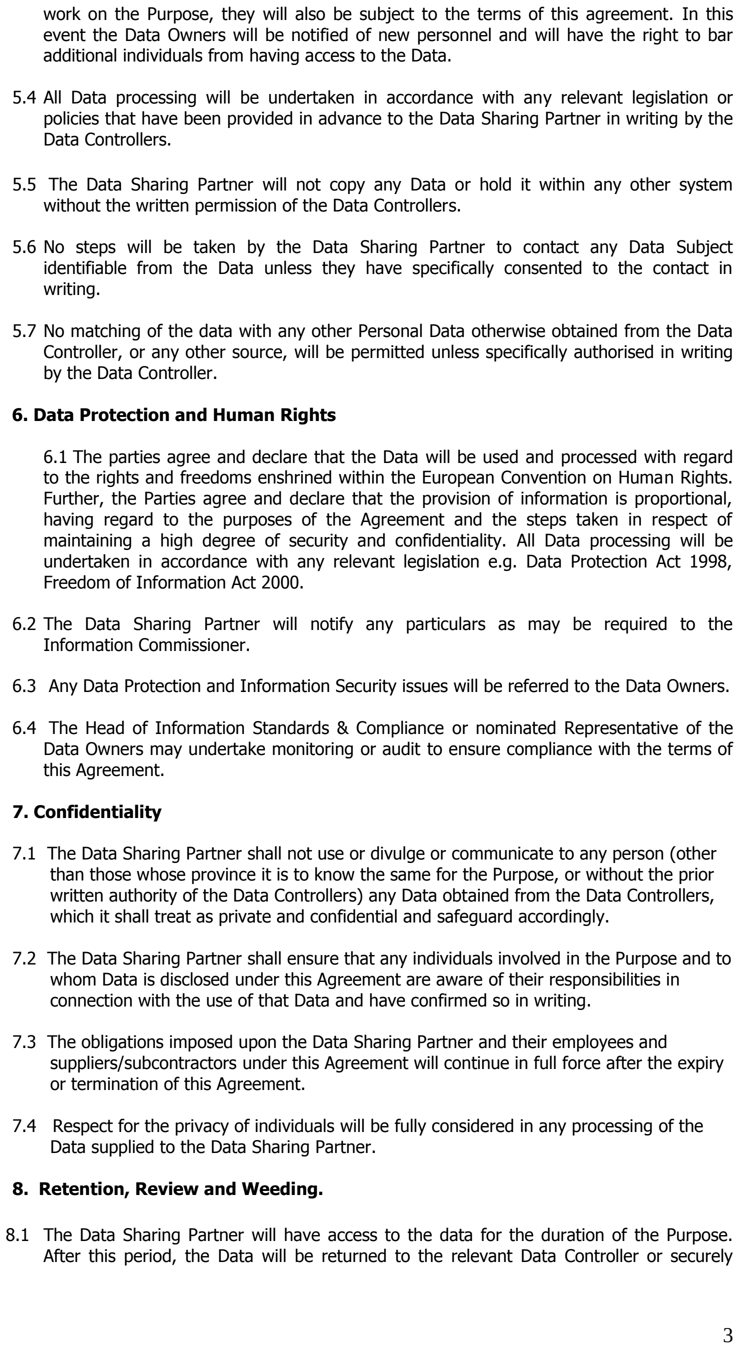
4.4 Ownership of the Data shall at all times remain with NBC.

5. Use, Disclosure and Publication

5.1 The Data will be used solely for the Purpose as defined above.

5.2 The Data Controllers will retain ownership of their respective Data. **No data may be copied, published, broadcast or otherwise disseminated to any third party** without the consent in writing of NBC.

5.3 Access to the Data will be restricted to the personnel listed in Appendix A who are the employees or Research Partners of the Data Sharing Partner. In the event that the Data Sharing Partner should assign or appoint any other employees or Research Partners to



work on the Purpose, they will also be subject to the terms of this agreement. In this event the Data Owners will be notified of new personnel and will have the right to bar additional individuals from having access to the Data.

5.4 All Data processing will be undertaken in accordance with any relevant legislation or policies that have been provided in advance to the Data Sharing Partner in writing by the Data Controllers.

5.5 The Data Sharing Partner will not copy any Data or hold it within any other system without the written permission of the Data Controllers.

5.6 No steps will be taken by the Data Sharing Partner to contact any Data Subject identifiable from the Data unless they have specifically consented to the contact in writing.

5.7 No matching of the data with any other Personal Data otherwise obtained from the Data Controller, or any other source, will be permitted unless specifically authorised in writing by the Data Controller.

6. Data Protection and Human Rights

6.1 The parties agree and declare that the Data will be used and processed with regard to the rights and freedoms enshrined within the European Convention on Human Rights. Further, the Parties agree and declare that the provision of information is proportional, having regard to the purposes of the Agreement and the steps taken in respect of maintaining a high degree of security and confidentiality. All Data processing will be undertaken in accordance with any relevant legislation e.g. Data Protection Act 1998, Freedom of Information Act 2000.

6.2 The Data Sharing Partner will notify any particulars as may be required to the Information Commissioner.

6.3 Any Data Protection and Information Security issues will be referred to the Data Owners.

6.4 The Head of Information Standards & Compliance or nominated Representative of the Data Owners may undertake monitoring or audit to ensure compliance with the terms of this Agreement.

7. Confidentiality

7.1 The Data Sharing Partner shall not use or divulge or communicate to any person (other than those whose province it is to know the same for the Purpose, or without the prior written authority of the Data Controllers) any Data obtained from the Data Controllers, which it shall treat as private and confidential and safeguard accordingly.

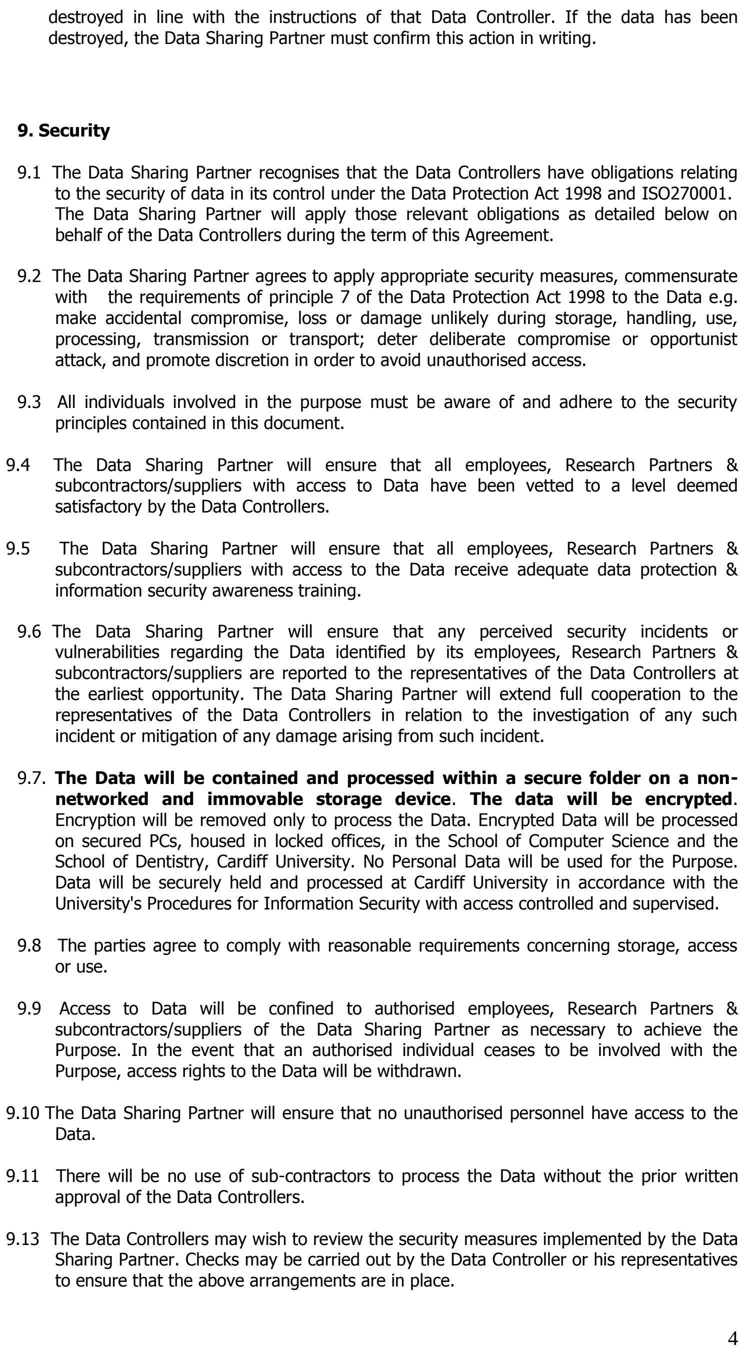
7.2 The Data Sharing Partner shall ensure that any individuals involved in the Purpose and to whom Data is disclosed under this Agreement are aware of their responsibilities in connection with the use of that Data and have confirmed so in writing.

7.3 The obligations imposed upon the Data Sharing Partner and their employees and suppliers/subcontractors under this Agreement will continue in full force after the expiry or termination of this Agreement.

7.4 Respect for the privacy of individuals will be fully considered in any processing of the Data supplied to the Data Sharing Partner.

8. Retention, Review and Weeding.

8.1 The Data Sharing Partner will have access to the data for the duration of the Purpose. After this period, the Data will be returned to the relevant Data Controller or securely



destroyed in line with the instructions of that Data Controller. If the data has been destroyed, the Data Sharing Partner must confirm this action in writing.

9. Security

9.1 The Data Sharing Partner recognises that the Data Controllers have obligations relating to the security of data in its control under the Data Protection Act 1998 and ISO270001. The Data Sharing Partner will apply those relevant obligations as detailed below on behalf of the Data Controllers during the term of this Agreement.

9.2 The Data Sharing Partner agrees to apply appropriate security measures, commensurate with the requirements of principle 7 of the Data Protection Act 1998 to the Data e.g. make accidental compromise, loss or damage unlikely during storage, handling, use, processing, transmission or transport; deter deliberate compromise or opportunist attack, and promote discretion in order to avoid unauthorised access.

9.3 All individuals involved in the purpose must be aware of and adhere to the security principles contained in this document.

9.4 The Data Sharing Partner will ensure that all employees, Research Partners & subcontractors/suppliers with access to Data have been vetted to a level deemed satisfactory by the Data Controllers.

9.5 The Data Sharing Partner will ensure that all employees, Research Partners & subcontractors/suppliers with access to the Data receive adequate data protection & information security awareness training.

9.6 The Data Sharing Partner will ensure that any perceived security incidents or vulnerabilities regarding the Data identified by its employees, Research Partners & subcontractors/suppliers are reported to the representatives of the Data Controllers at the earliest opportunity. The Data Sharing Partner will extend full cooperation to the representatives of the Data Controllers in relation to the investigation of any such incident or mitigation of any damage arising from such incident.

9.7. **The Data will be contained and processed within a secure folder on a non-networked and immovable storage device**. **The data will be encrypted**. Encryption will be removed only to process the Data. Encrypted Data will be processed on secured PCs, housed in locked offices, in the School of Computer Science and the School of Dentistry, Cardiff University. No Personal Data will be used for the Purpose. Data will be securely held and processed at Cardiff University in accordance with the University's Procedures for Information Security with access controlled and supervised.

9.8 The parties agree to comply with reasonable requirements concerning storage, access or use.

9.9 Access to Data will be confined to authorised employees, Research Partners & subcontractors/suppliers of the Data Sharing Partner as necessary to achieve the Purpose. In the event that an authorised individual ceases to be involved with the Purpose, access rights to the Data will be withdrawn.

9.10 The Data Sharing Partner will ensure that no unauthorised personnel have access to the Data.

9.11 There will be no use of sub-contractors to process the Data without the prior written approval of the Data Controllers.

9.13 The Data Controllers may wish to review the security measures implemented by the Data Sharing Partner. Checks may be carried out by the Data Controller or his representatives to ensure that the above arrangements are in place.

9.14The Data Sharing Partner will ensure that the Personal Data accessed is not used other than as identified within this Agreement, and that the Agreement is complied with.

10. Subject Access Requests

10.1The Data Protection Act 1998, s.7 provides individuals with a right to have access to data, which is held about them, by a Data Controller, on computer or manual files – unless an exemption applies where information can be withheld under certain circumstances.

10.2 The Data Sharing Partner shall give all reasonable assistance as is necessary to the Data Controllers in order to enable them to:

- Comply with request for subject access from the data subjects;
- Respond to Information Notices (as defined by the Data Protection Act 1998 served upon him by the Information Commissioner;
- Respond to complaints from data subjects;
- Investigate any breach or alleged breach of the Act.

in accordance with its statutory obligations under the Data Protection Act 1998.

10.3 The receipt by the Data Sharing Partner of any subject access requests, to the Data covered by this Agreement must be reported and any relevant information be forwarded, at the earliest opportunity to the Data Protection and Information Security Section who will arrange the relevant response to that request.

10.4 This Agreement also acts in fulfilment of part of the responsibilities of the Data Controllers as required by paragraphs 11 and 12 of Schedule 1, Part II of the Data Protection Act 1998.

11. Complaints & Breaches

11.1 Any complaints in respect of the Data and the processing thereof, will be brought to the attention of the designated manager if Cardiff University and will be dealt with in accordance with the Cardiff University's internal complaints procedure.

11.2 The parties agree that any breach of this Data Sharing Agreement will seriously undermine and affect the credibility of the Purpose, and may render parties liable for breach of the law.

12. Disputes

13.1 Any disputes between the parties arising out of or in respect of this Agreement will be decided in accordance with the laws of England & Wales and will be subject to the jurisdiction of the courts of England & Wales.

14. Miscellaneous

14.1 This Agreement will continue throughout the term of the Purpose and for as long as the Data is held by the Data Sharing Partner unless superseded or replaced by the agreement of the parties.

14.2 To ensure the terms of this Agreement are being adhered to, the Data Controllers and Data Sharing Partner will each delegate a named individual to oversee this function. In the event that the individual(s) cease to continue in their roles, replacements must be identified.

14.3 The Agreement will be terminated by any party by 60 days' written notice to the other parties of any such termination.

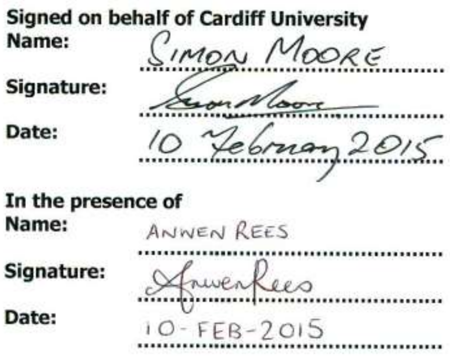
14.4 This Agreement may be executed in any number of counterparts, each of which shall be deemed to be an original, and all of which taken together shall constitute one and the same instrument. A PDF copy of a signature shall constitute an original signature for all purposes.

Signed on behalf of Northants Police
Name: ....
Signature: ....
Date: ....

In the presence of
Name: ....
Signature: ....
Date: ....

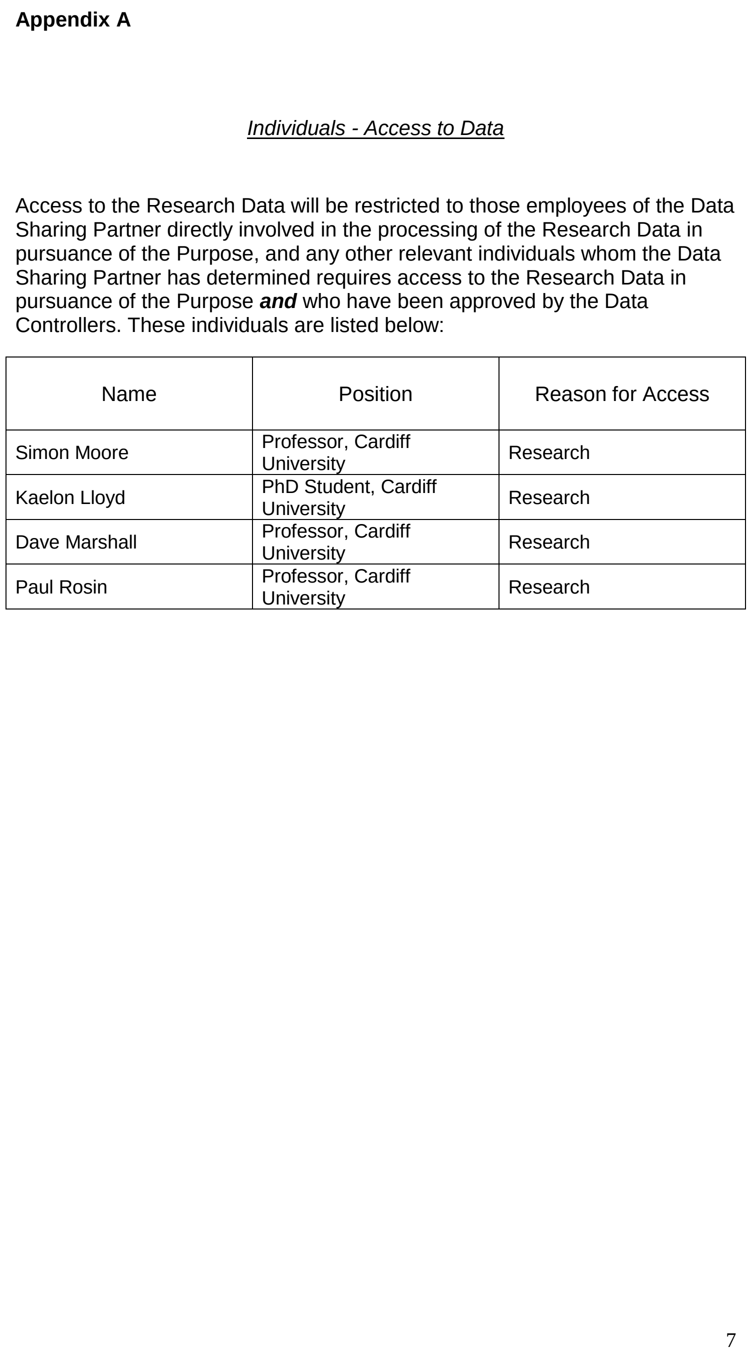
Signed on behalf of Northampton Borough Council
Name: ....
Signature: ....
Date: ....

In the presence of
Name: ....
Signature: ....
Date: ....



Signed on behalf of Cardiff University
Name: SIMON MOORE
Signature: Simon Moore
Date: 10 February 2015

In the presence of
Name: ANWEN REES
Signature: Anwen Rees
Date: 10-FEB-2015

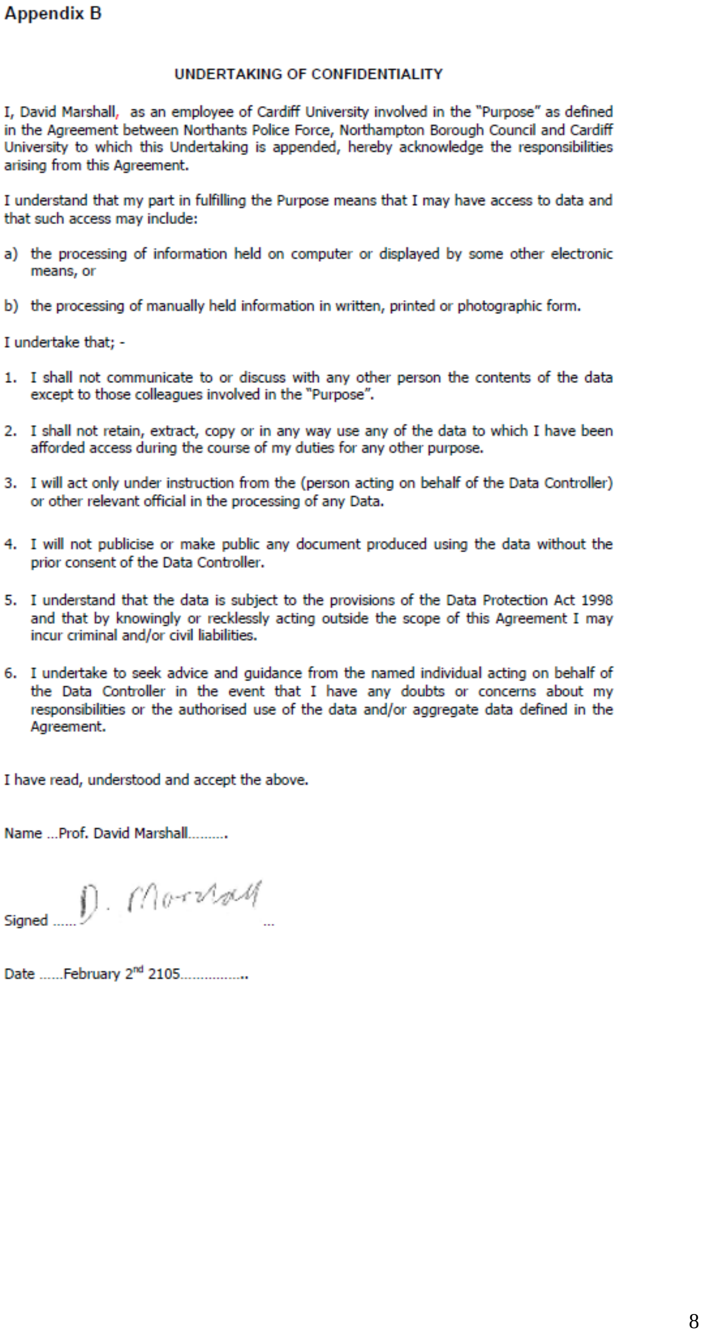


Appendix A

Individuals - Access to Data

Access to the Research Data will be restricted to those employees of the Data Sharing Partner directly involved in the processing of the Research Data in pursuance of the Purpose, and any other relevant individuals whom the Data Sharing Partner has determined requires access to the Research Data in pursuance of the Purpose ***and*** who have been approved by the Data Controllers. These individuals are listed below:

Name	Position	Reason for Access
Simon Moore	Professor, Cardiff University	Research
Kaelon Lloyd	PhD Student, Cardiff University	Research
Dave Marshall	Professor, Cardiff University	Research
Paul Rosin	Professor, Cardiff University	Research



Appendix B

UNDERTAKING OF CONFIDENTIALITY

I, David Marshall, as an employee of Cardiff University involved in the "Purpose" as defined in the Agreement between Northants Police Force, Northampton Borough Council and Cardiff University to which this Undertaking is appended, hereby acknowledge the responsibilities arising from this Agreement.

I understand that my part in fulfilling the Purpose means that I may have access to data and that such access may include:

a) the processing of information held on computer or displayed by some other electronic means, or

b) the processing of manually held information in written, printed or photographic form.

I undertake that; -

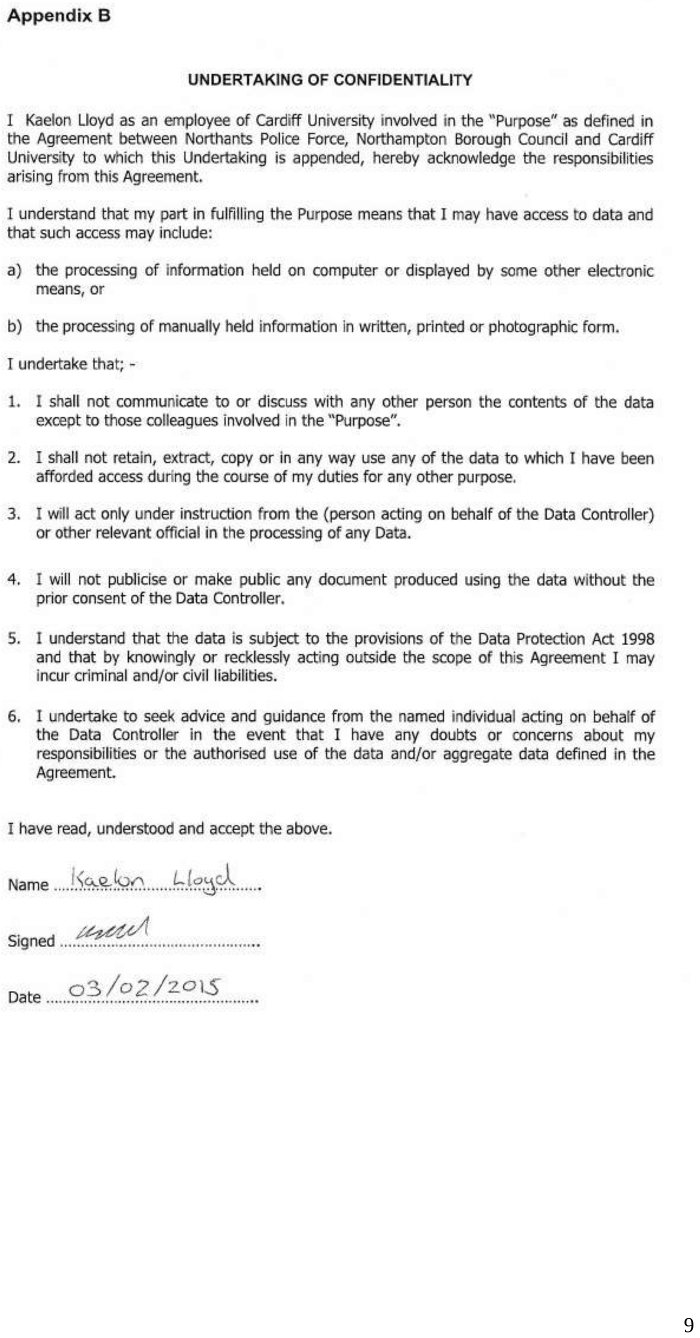
1. I shall not communicate to or discuss with any other person the contents of the data except to those colleagues involved in the "Purpose".
2. I shall not retain, extract, copy or in any way use any of the data to which I have been afforded access during the course of my duties for any other purpose.
3. I will act only under instruction from the (person acting on behalf of the Data Controller) or other relevant official in the processing of any Data.
4. I will not publicise or make public any document produced using the data without the prior consent of the Data Controller.
5. I understand that the data is subject to the provisions of the Data Protection Act 1998 and that by knowingly or recklessly acting outside the scope of this Agreement I may incur criminal and/or civil liabilities.
6. I undertake to seek advice and guidance from the named individual acting on behalf of the Data Controller in the event that I have any doubts or concerns about my responsibilities or the authorised use of the data and/or aggregate data defined in the Agreement.

I have read, understood and accept the above.

Name ...Prof. David Marshall.........

Signed ...... D. Marshall ...

Date ......February 2nd 2105................



Appendix B

UNDERTAKING OF CONFIDENTIALITY

I Kaelon Lloyd as an employee of Cardiff University involved in the "Purpose" as defined in the Agreement between Northants Police Force, Northampton Borough Council and Cardiff University to which this Undertaking is appended, hereby acknowledge the responsibilities arising from this Agreement.

I understand that my part in fulfilling the Purpose means that I may have access to data and that such access may include:

a) the processing of information held on computer or displayed by some other electronic means, or

b) the processing of manually held information in written, printed or photographic form.

I undertake that; -

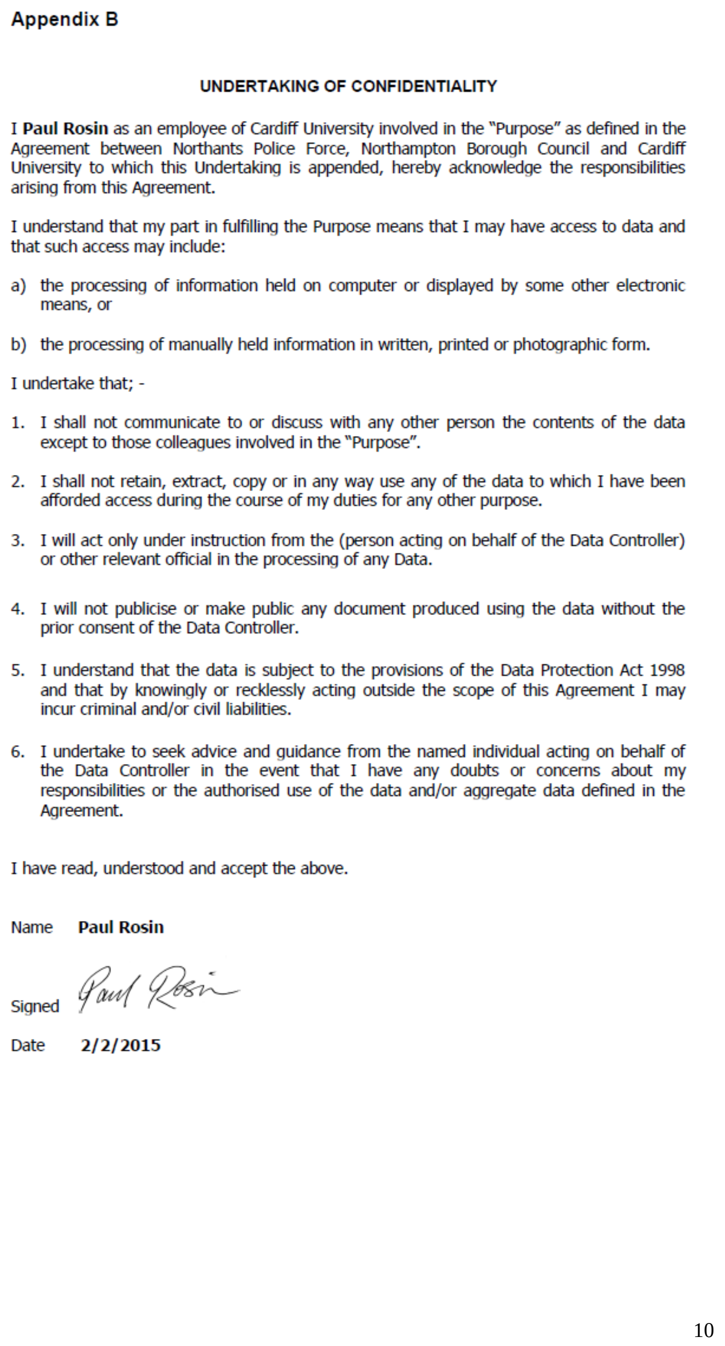
1. I shall not communicate to or discuss with any other person the contents of the data except to those colleagues involved in the "Purpose".
2. I shall not retain, extract, copy or in any way use any of the data to which I have been afforded access during the course of my duties for any other purpose.
3. I will act only under instruction from the (person acting on behalf of the Data Controller) or other relevant official in the processing of any Data.
4. I will not publicise or make public any document produced using the data without the prior consent of the Data Controller.
5. I understand that the data is subject to the provisions of the Data Protection Act 1998 and that by knowingly or recklessly acting outside the scope of this Agreement I may incur criminal and/or civil liabilities.
6. I undertake to seek advice and guidance from the named individual acting on behalf of the Data Controller in the event that I have any doubts or concerns about my responsibilities or the authorised use of the data and/or aggregate data defined in the Agreement.

I have read, understood and accept the above.

Name ....Kaelon Lloyd.....

Signed ....

Date ...03/02/2015.........



Appendix B

UNDERTAKING OF CONFIDENTIALITY

I **Paul Rosin** as an employee of Cardiff University involved in the "Purpose" as defined in the Agreement between Northants Police Force, Northampton Borough Council and Cardiff University to which this Undertaking is appended, hereby acknowledge the responsibilities arising from this Agreement.

I understand that my part in fulfilling the Purpose means that I may have access to data and that such access may include:

a) the processing of information held on computer or displayed by some other electronic means, or

b) the processing of manually held information in written, printed or photographic form.

I undertake that; -

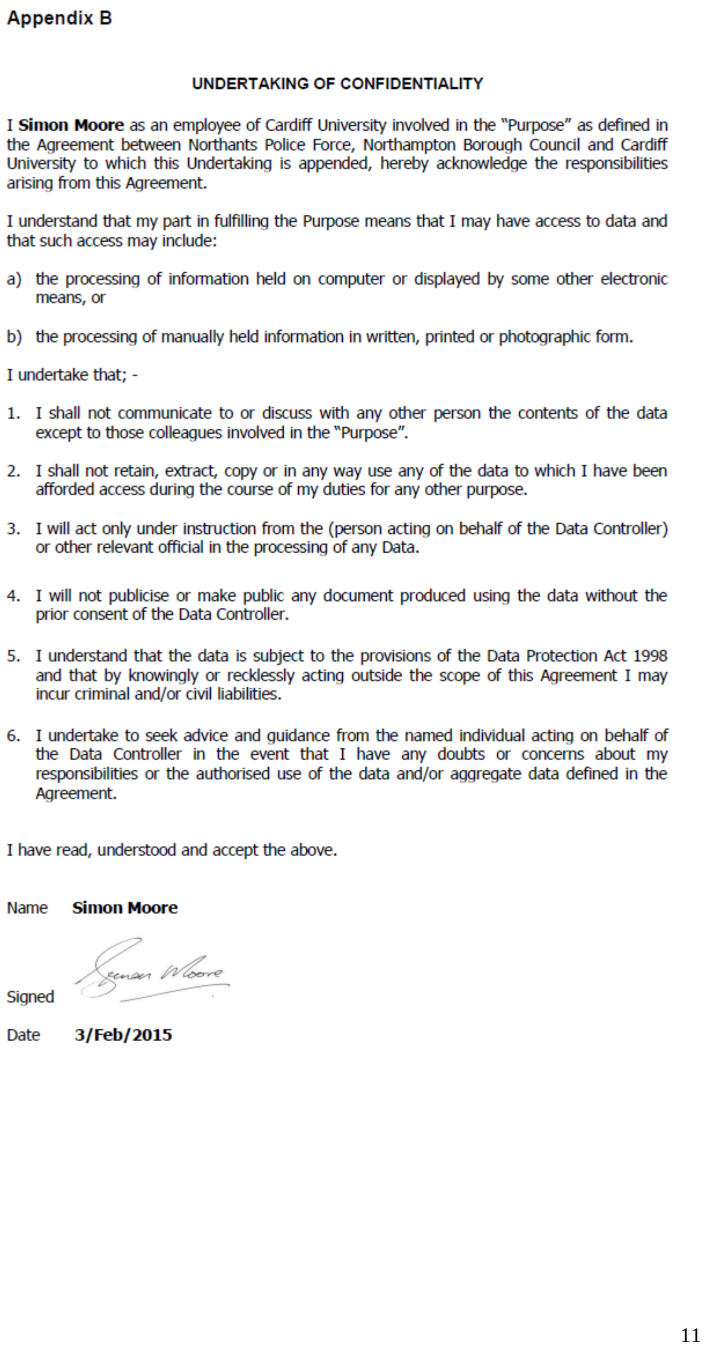
1. I shall not communicate to or discuss with any other person the contents of the data except to those colleagues involved in the "Purpose".
2. I shall not retain, extract, copy or in any way use any of the data to which I have been afforded access during the course of my duties for any other purpose.
3. I will act only under instruction from the (person acting on behalf of the Data Controller) or other relevant official in the processing of any Data.
4. I will not publicise or make public any document produced using the data without the prior consent of the Data Controller.
5. I understand that the data is subject to the provisions of the Data Protection Act 1998 and that by knowingly or recklessly acting outside the scope of this Agreement I may incur criminal and/or civil liabilities.
6. I undertake to seek advice and guidance from the named individual acting on behalf of the Data Controller in the event that I have any doubts or concerns about my responsibilities or the authorised use of the data and/or aggregate data defined in the Agreement.

I have read, understood and accept the above.

Name **Paul Rosin**

Signed Paul Rosin

Date **2/2/2015**



Appendix B

UNDERTAKING OF CONFIDENTIALITY

I **Simon Moore** as an employee of Cardiff University involved in the "Purpose" as defined in the Agreement between Northants Police Force, Northampton Borough Council and Cardiff University to which this Undertaking is appended, hereby acknowledge the responsibilities arising from this Agreement.

I understand that my part in fulfilling the Purpose means that I may have access to data and that such access may include:

a) the processing of information held on computer or displayed by some other electronic means, or

b) the processing of manually held information in written, printed or photographic form.

I undertake that; -

1. I shall not communicate to or discuss with any other person the contents of the data except to those colleagues involved in the "Purpose".
2. I shall not retain, extract, copy or in any way use any of the data to which I have been afforded access during the course of my duties for any other purpose.
3. I will act only under instruction from the (person acting on behalf of the Data Controller) or other relevant official in the processing of any Data.
4. I will not publicise or make public any document produced using the data without the prior consent of the Data Controller.
5. I understand that the data is subject to the provisions of the Data Protection Act 1998 and that by knowingly or recklessly acting outside the scope of this Agreement I may incur criminal and/or civil liabilities.
6. I undertake to seek advice and guidance from the named individual acting on behalf of the Data Controller in the event that I have any doubts or concerns about my responsibilities or the authorised use of the data and/or aggregate data defined in the Agreement.

I have read, understood and accept the above.

Name **Simon Moore**

Signed Simon Moore

Date **3/Feb/2015**

Appendix B

Quality: Full-Reference and No-Reference correlation

The correlation between full-reference and no-reference image quality assessment methods.

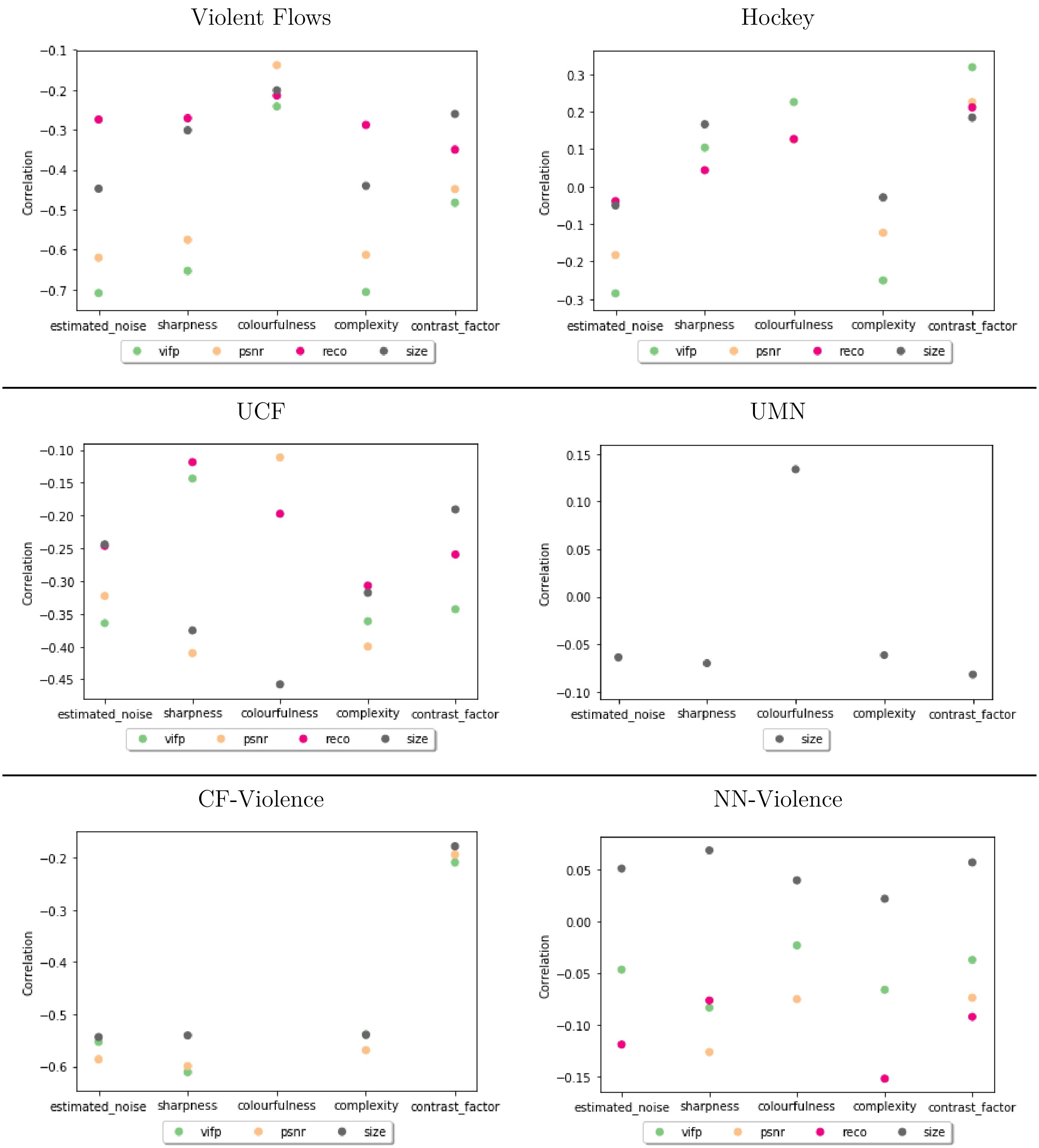


Table B.1: Statistically significant (< 0.05) Point Biserial correlation coefficient between blind quality measurements and binary prediction using reference based quality analysis..

Appendix C

No Reference Quality Correlation

Correlation between each no-reference image quality assessment methods.

	colourfulness	complexity	contrast_factor	estimated_noise	sharpness
colourfulness	1.000000	0.077423	0.203425	-0.022511	0.247942
complexity	0.077423	1.000000	0.560339	0.702943	0.574084
contrast_factor	0.203425	0.560339	1.000000	0.219389	0.484698
estimated_noise	-0.022511	0.702943	0.219389	1.000000	0.345738
sharpness	0.247942	0.574084	0.484698	0.345738	1.000000

Table C.1: NN-Violence Blind Quality Correlation

	colourfulness	complexity	contrast_factor	estimated_noise	sharpness
colourfulness	1.000000	-0.714089	-0.758384	-0.649331	-0.607384
complexity	-0.714089	1.000000	0.927929	0.981650	0.965552
contrast_factor	-0.758384	0.927929	1.000000	0.902728	0.895552
estimated_noise	-0.649331	0.981650	0.902728	1.000000	0.952375
sharpness	-0.607384	0.965552	0.895552	0.952375	1.000000

Table C.2: UMN Blind Quality Correlation

	colourfulness	complexity	contrast_factor	estimated_noise	sharpness
colourfulness	1.000000	0.217836	0.159848	0.160228	0.019311
complexity	0.217836	1.000000	0.221632	0.951859	0.295015
contrast_factor	0.159848	0.221632	1.000000	0.135176	0.358787
estimated_noise	0.160228	0.951859	0.135176	1.000000	0.311157
sharpness	0.019311	0.295015	0.358787	0.311157	1.000000

Table C.3: Hockey Blind Quality Correlation

	colourfulness	complexity	contrast_factor	estimated_noise	sharpness
colourfulness	1.000000	0.366699	-0.026336	0.339385	-0.089466
complexity	0.366699	1.000000	0.415317	0.956623	0.321041
contrast_factor	-0.026336	0.415317	1.000000	0.302220	0.486171
estimated_noise	0.339385	0.956623	0.302220	1.000000	0.210222
sharpness	-0.089466	0.321041	0.486171	0.210222	1.000000

Table C.4: UCF Blind Quality Correlation

	colourfulness	complexity	contrast_factor	estimated_noise	sharpness
colourfulness	1.000000	0.178709	0.262209	0.180310	0.209685
complexity	0.178709	1.000000	0.578366	0.963129	0.630525
contrast_factor	0.262209	0.578366	1.000000	0.542525	0.599104
estimated_noise	0.180310	0.963129	0.542525	1.000000	0.666680
sharpness	0.209685	0.630525	0.599104	0.666680	1.000000

Table C.5: Violent Flows Blind Quality Correlation

	colourfulness	complexity	contrast_factor	estimated_noise	sharpness
colourfulness	1.000000	0.265929	-0.141347	0.247825	-0.082208
complexity	0.265929	1.000000	0.557140	0.954916	0.529659
contrast_factor	-0.141347	0.557140	1.000000	0.538375	0.583404
estimated_noise	0.247825	0.954916	0.538375	1.000000	0.584906
sharpness	-0.082208	0.529659	0.583404	0.584906	1.000000

Table C.6: CF-Violence Blind Quality Correlation

Appendix D

Agent Based Modelling: Bivariate scatter plots

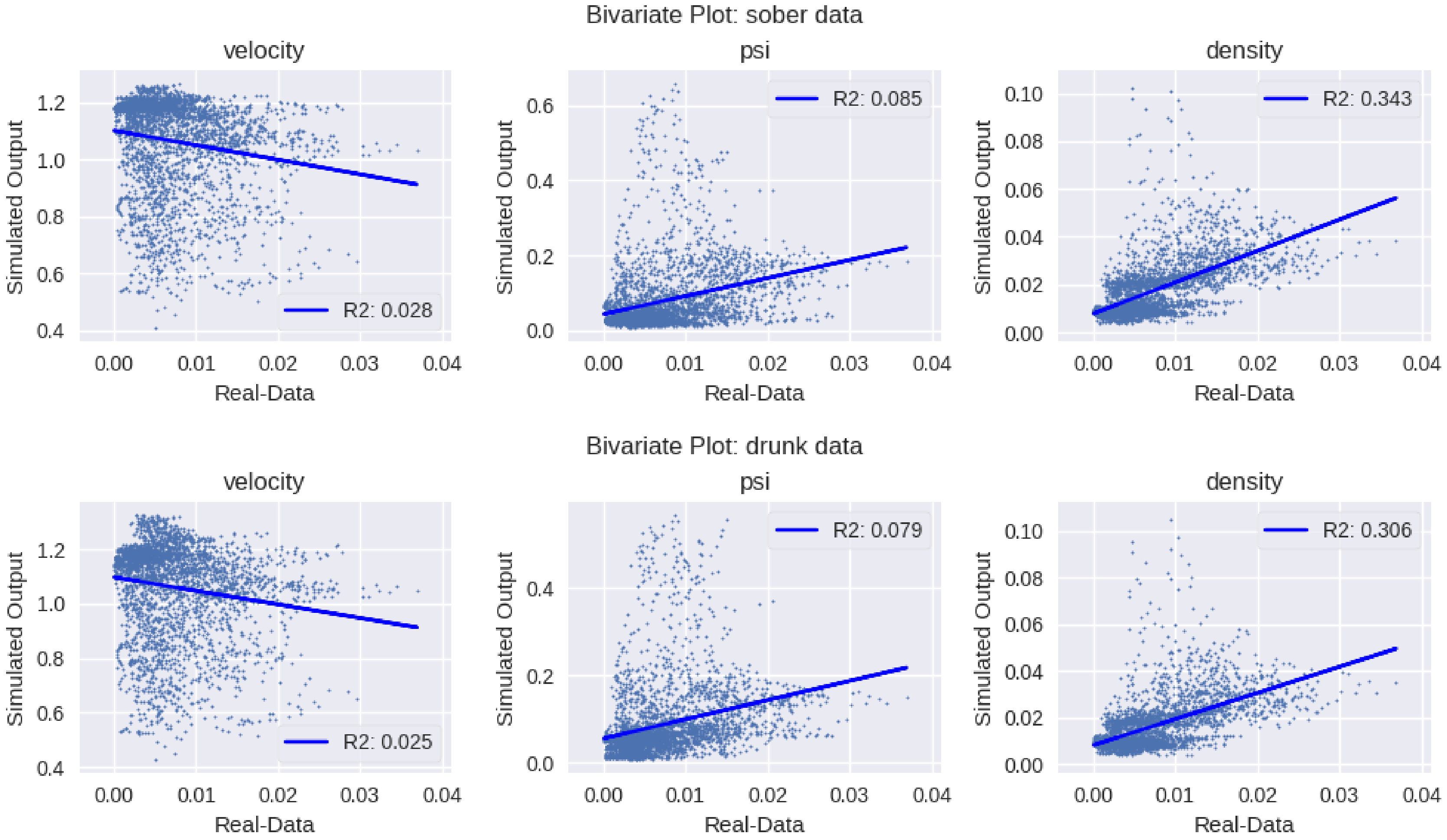


Figure D.1: Bivariate scatter plots between simulation output and violent crime in Northampton. Top: Non-intoxicate, Bottom: Intoxicated.

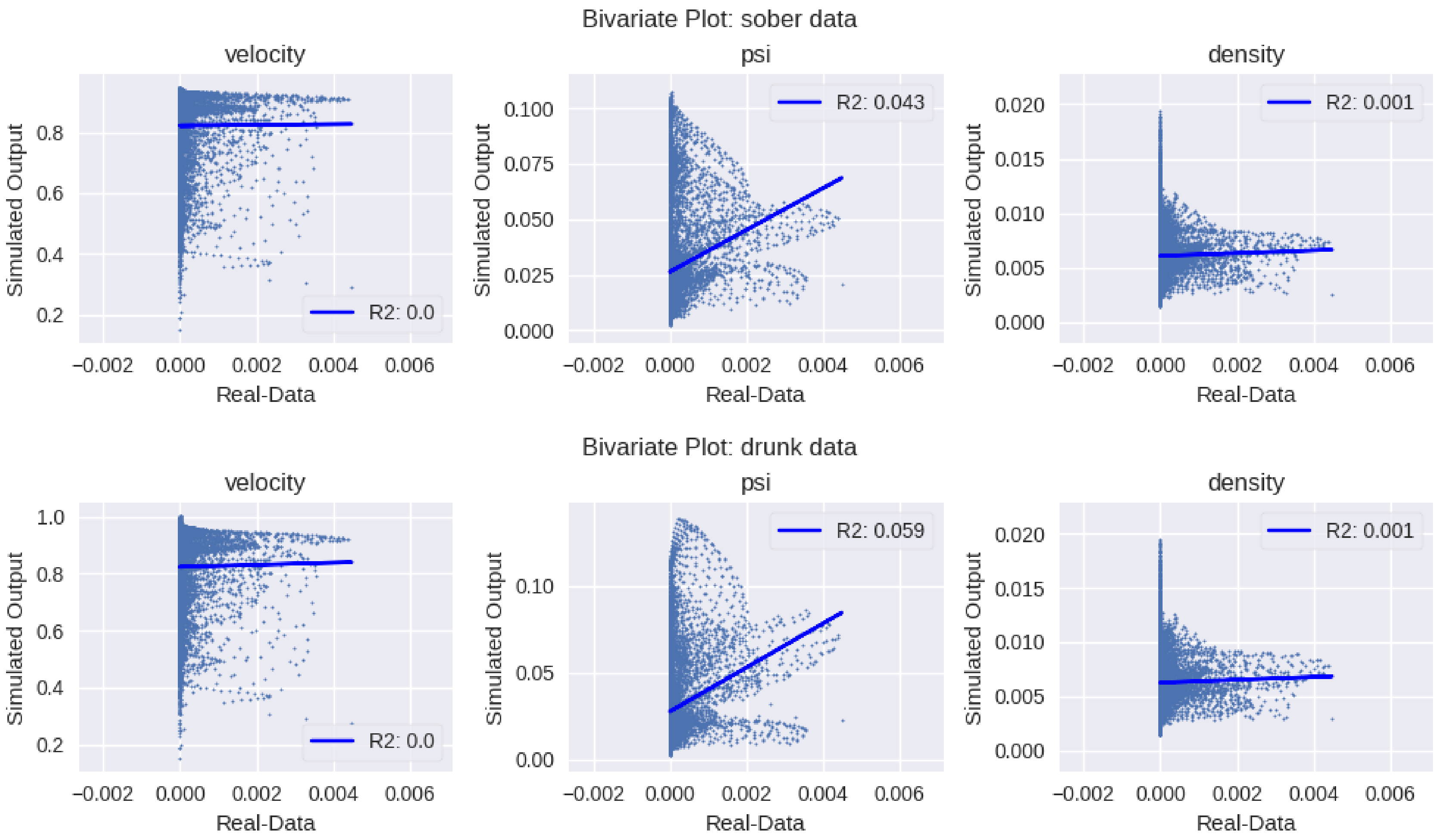


Figure D.2: Bivariate scatter plots between simulation output and violent crime in Cardiff. Top: Non-intoxicate, Bottom: Intoxicated.

Bibliography

[1] Christina Aiello and Emmanuel Agu. Investigating postural sway features, normalization and personalization in detecting blood alcohol levels of smartphone users. In *Wireless Health*, pages 73–80, 2016.

[2] Khosro Bahrami and Alex C Kot. A fast approach for no-reference image sharpness assessment based on maximum local variation. *IEEE Signal Processing Letters*, 21(6):751–755, 2014.

[3] Ankan Bansal and K. S. Venkatesh. People Counting in High Density Crowds from Still Images. *International Journal of Computer and Electrical Engineering*, 7(5):316–324, 2015.

[4] Ola W. Barnett, Ronald W. Fagan, and Jolyne M. Booker. Hostility and stress as mediators of aggression in violent men. *Journal of Family Violence*, 6(3):217–241, 1991.

[5] Piotr Bilinski and Francois Bremond. Human violence recognition and detection in surveillance videos. In *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 30–36. IEEE, 2016.

[6] Soma Biswas and Vikas Gupta. Abnormality detection in crowd videos by tracking sparse components. *Machine Vision and Applications*, 28(1-2):35–48, 2016.

[7] Victor J. Blue and Jeffrey L. Adler. Cellular automata microsimulation for modeling bi-directional pedestrian walkways. *Transportation Research Part B: Methodological*, 35(3):293–312, 2001.

[8] Scott Blunsden and RB Fisher. The behave video dataset: ground truthed video for multi-person behavior classification.

[9] Anthony A. Braga, Andrew V. Papachristos, and David M. Hureau. The effects of hot spots policing on crime: An updated systematic review and meta-analysis. *Justice quarterly*, 31(4):633–663, 2014.

[10] Anthony A. Braga, David L. Weisburd, Elin J. Waring, Lorraine Green Mazerolle, William Spelman, and Francis Gajewski. Problem-oriented policing in violent crime places: A randomized controlled experiment. *Criminology*, 37(3):541–580, 1999.

[11] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, oct 2001.

[12] Francois Bremond, Nicolas Maillot, Monique Thonnat, and Van-Thinh Vu. *Ontologies for video events*. PhD thesis, INRIA, 2004.

[13] François Brémond, Monique Thonnat, and Marcos Zúniga. Video-understanding framework for automatic behavior recognition. *Behavior Research Methods*, 38(3):416–426, 2006.

[14] Vince D. Calhoun, James J. Pekar, and Godfrey D. Pearlson. Alcohol intoxication effects on simulated driving: exploring alcohol-dose effects on brain activation using functional MRI. *Neuropsychopharmacology*, 29(11):2097, 2004.

[15] SP Chainey. Examining the influence of cell size and bandwidth size on kernel density estimation crime hotspot maps for predicting spatial patterns of crime. *Bulletin of the Geographical Society of Liege*, 60:7–19, 2013.

[16] Antoni B. Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–7. IEEE, 2008.

[17] Chunyu Chen, Yu Shao, and Xiaojun Bi. Detection of Anomalous Crowd Behavior Based on the Acceleration Feature. *IEEE Sensors Journal*, 15(12):7252–7261, 2015.

[18] Mohcine Chraibi and Jun Zhang. JuPedSim: an open framework for simulating and analyzing the dynamics of pedestrians. In *SUMO Conference 2016*, volume 30, pages 127–134. Jülich Supercomputing Center, 2016.

[19] Mei Ling Chu, Paolo Parigi, Jean-Claude Latombe, and Kincho H. Law. Simulating effects of signage, groups, and crowds on emergent evacuation patterns. *Ai & Society*, 30(4):493–507, 2015.

[20] Sheldon Cohen. Aftereffects of stress on human performance and social behavior: a review of research and theory. *Psychological bulletin*, 88(1):82, 1980.

[21] D. Conte, P. Foggia, G. Percannella, F. Tufano, and M. Vento. Counting moving people in videos by salient points detection. *Proceedings - International Conference on Pattern Recognition*, pages 1743–1746, 2010.

[22] Ian W. Craig. The importance of stress and genetic variation in human aggression. *Bioessays*, 29(3):227–236, 2007.

[23] Michael Cusimano, Sean Marshall, Claus Rinner, Depeng Jiang, and Mary Chipman. Patterns of urban violent injury: a spatio-temporal analysis. *PLoS One*, 5(1):e8669, 2010.

[24] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005.*

IEEE Computer Society Conference on, volume 1, pages 886–893. IEEE, 2005.

[25] Purves Dale, J. Augustine George, Fitzpatrick David, C. Katz Lawrence, LaMantia Anthony-Samuel, McNamara O. James, and Mark S. Williams. *Neuroscience*. Sinauer Associates, 2001.

[26] a. Datta, M. Shah, and N. Da Vitoria Lobo. Person-on-person violence detection in video data. In *Object recognition supported by user interaction for service robots*, volume 1, pages 433–438. IEEE Comput. Soc, 2002.

[27] Richard J. Dawson, Roger Peppe, and Miao Wang. An agent-based model for risk-based flood incident management. *Natural hazards*, 59(1):167–189, 2011.

[28] Shinichi Demura and Masanobu Uchiyama. Influence of moderate alcohol ingestion on gait. *Sport Sciences for Health*, 4(1):21–26, 2008.

[29] O. Deniz, I. Serrano, G. Bueno, and T. K. Kim. Fast violence detection in video. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 2, pages 478–485, Jan 2014.

[30] C. Diffley and E. Wallace. CCTV: Making it work. Technical report, Great Britain, Home Office, Police Scientific Development Branch, 1998.

[31] Kunio Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics*, 31(4-5):198–211, 2007.

[32] Zhihong Dong, Jie Qin, and Yunhong Wang. Multi-stream Deep Networks for Person to Person Violence Detection in Videos. In Tieniu Tan, Xuelong Li, Xilin Chen, Jie Zhou, Jian Yang, and Hong Cheng, editors, *Pattern Recognition*, pages 517–531, Singapore, 2016. Springer Singapore.

[33] Anne Dray, Lorraine Mazerolle, Pascal Perez, and Alison Ritter. Policing Australia's 'heroin drought': using an agent-based model to simulate alternative outcomes. *Journal of Experimental Criminology*, 4(3):267–287, 2008.

[34] Richard Dubourg, Joe Hamed, and Jamie Thorns. The economic and social costs of crime against individuals and households 2003/04. *London: Home Office, Economics and Resource Analysis Research, Development and Statistics*, 2005.

[35] John Eck, Spencer Chainey, James Cameron, and Ronald Wilson. Mapping crime: Understanding hotspots. *National Institute of Justice*, pages 1–71, 2005.

[36] John Flatley. Crime in england and wales: year ending june 2017. *Office for National Statistics*, 2017.

[37] Curtis Florence, Jonathan Shepherd, Iain Brennan, and Thomas Simon. Effectiveness of anonymised information sharing and use in health service, police, and local government partnership for preventing violence related injury: experimental study and time series analysis. *BMJ*, 342, 2011.

[38] Curtis Florence, Jonathan Shepherd, Iain Brennan, and Thomas R. Simon. An economic evaluation of anonymised information sharing in a partnership between health services, police and local government for preventing violence-related injury. *Injury prevention*, 20(2):108–114, 2014.

[39] Audit Commission for Local Authorities, the National Health Service in England, and Wales. *Helping with enquiries: Tackling crime effectively.* HM Stationery Office, 1993.

[40] Yuan Gao, Hong Liu, Xiaohu Sun, Can Wang, and Yi Liu. Violence detection using Oriented VIolent Flows. *Image and Vision Computing*, 48-49(2015):37–41, 2015.

[41] Graeme Gerrard, Garry Parkins, Ian Cunningham, Wayne Jones, Samantha Hill, and Sarah Douglas. National CCTV strategy. *UK Home Office*, 2007.

[42] Graeme Gerrard and Richard Thompson. Two million cameras in the UK. *CCTVImage*, 42(42):10–12, 2011.

[43] Martin Gill and Angela Spriggs. *Assessing the impact of CCTV.* Home Office Research, Development and Statistics Directorate London, 2005.

[44] Gerhard Gmel, John Holmes, and Joseph Studer. Are alcohol outlet densities strongly associated with alcohol-related outcomes? A critical review of recent evidence. *Drug and alcohol review*, 35(1):40–54, 2016.

[45] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* MIT Press, 2016. `http://www.deeplearningbook.org`.

[46] M. Goodwin, C. Johnstone, and K. Williams. New Spaces of Law Enforcement: Closed Circuit Television, Public Behaviour, and the Policing of Public Space. In *Association of American Geographers Annual Conference, Boston, MA, March*, 1998.

[47] Benjamin Jervis Goold. *CCTV and policing: Public area surveillance and police practices in Britain.* Oxford University Press on Demand, 2004.

[48] Wilpen L Gorr and YongJei Lee. Longitudinal study of crime hot spots: dynamics and impact on part 1 violent crime. In *Proceedings of the 32nd international symposium on forecasting*, 2012.

[49] Ismael Serrano Gracia, Oscar Deniz Suarez, Gloria Bueno Garcia, and Tae Kyun Kim. Fast fight detection. *PLoS ONE*, 10(4):1–19, 2015.

[50] Benjamin J. Gray, Emma R. Barton, Alisha R. Davies, Sara J. Long, Janine Roderick, and Mark A. Bellis. A shared data approach more accurately represents the rates and patterns of violence with injury assaults. *J Epidemiol Community Health*, 71(12):1218–1224, 2017.

[51] Mary Wilson Green. The appropriate and effective use of security technologies in US schools: a guide for schools and law enforcement agencies. Technical report, Sandia National Laboratories, 2005.

[52] Edward T. Hall. A system for the notation of proxemic behavior. *American anthropologist*, 65(5):1003–1026, 1963.

[53] R. M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural Features for Image Classification. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3(6):610–621, November 1973.

[54] Robert M Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979.

[55] David Hasler and Sabine Süsstrunk. Measuring Colourfulness in Natural Images. *Proc. IS&amp;T/SPIE Electronic Imaging 2003: Human Vision and Electronic Imaging VIII*, 5007:87–95, 2003.

[56] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6, 2012.

[57] Glenn I Hawe, Graham Coates, Duncan T Wilson, and Roger S Crouch. Agent-based simulation of emergency response to plan the allocation of resources for a hypothetical two-site major incident. *Engineering Applications of Artificial Intelligence*, 46:336–345, 2015.

[58] Leslie A Hayduk. Personal space: Where we now stand. *Psychological bulletin*, 94(2):293, 1983.

[59] Dirk Helbing, Anders Johansson, and Habib Zein Al-Abideen. Dynamics of crowd disasters: An empirical study. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 75(4), 2007.

[60] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995.

[61] Dirk Helbing, Péter Molnár, Illés J Farkas, and Kai Bolay. Self-organizing pedestrian movement. *Environment and planning B: planning and design*, 28(3):361–383, 2001.

[62] Lily Hirsch, Kirrilly Thompson, and Danielle Every. Frustrations, Fights, and Friendships: The Physical, Emotional, and Behavioural Effects of High-Density Crowding on Mumbai's Suburban Rail Passengers. *The Qualitative Report*, 22(2):550–566, 2017.

[63] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[64] Christina J. Howard, Iain D. Gilchrist, Tom Troscianko, Ardhendu Behera, and David C. Hogg. Task relevance predicts gaze in videos of real moving scenes. *Experimental brain research*, 214(1):131, 2011.

[65] Christina Jayne Howard, Tom Troscianko, Iain D. Gilchrist, Ardhendu Behera, and David C. Hogg. Suspiciousness perception in dynamic scenes: a comparison of CCTV operators and novices. *Frontiers in human neuroscience*, 7:441, 2013.

[66] John Immerkaer. Fast noise variance estimation. *Computer vision and image understanding*, 64(2):300–302, 1996.

[67] Uday Jain. Effects of population density and resources on the feeling of crowding and personal space. *The Journal of social psychology*, 127(3):331–338, 1987.

[68] E. C. Jansen, H. H. Thyssen, and J. Brynskov. Gait analysis after intake of increasing amounts of alcohol. *International Journal of Legal Medicine*, 94(2):103–107, 1985.

[69] Kim R. Kanaga and Mark Flynn. The relationship between invasion of personal space and stress. *Human Relations*, 34(3):239–248, 1981.

[70] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[71] George L. Kelling, Tony Pate, Duane Dieckman, and Charles E. Brown. The Kansas city preventive patrol experiment. *Police Foundation*, 1974.

[72] Louis Kratz and Ko Nishino. Tracking pedestrians using local spatio-temporal motion patterns in extremely crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):987–1002, 2012.

[73] Menno R. Kruk, Jozsef Halasz, Wout Meelis, and Jozsef Haller. Fast positive feedback between the adrenocortical stress response and a brain mechanism involved in aggressive behavior. *Behavioral neuroscience*, 118(5):1062, 2004.

[74] William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[75] T. L. Lai, Herbert Robbins, and C. Zi Wei. Strong consistency of least squares estimates in multiple regression II. *Journal of Multivariate Analysis*, 9(3):343–361, 1979.

[76] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016.

[77] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *26th IEEE Conference on*

Computer Vision and Pattern Recognition, CVPR, pages 1–8. IEEE, jun 2008.

[78] Marco Leo, G. Medioni, M. Trivedi, Takeo Kanade, and Giovanni Maria Farinella. Computer vision for assistive technologies. *Computer Vision and Image Understanding*, 154:1–15, 2017.

[79] Longzhen Li, Tahir Nawaz, and James Ferryman. Pets 2015: datasets and challenge. In *Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on*, pages 1–6. IEEE, 2015.

[80] Xin Li. Blind image quality assessment. In *Proceedings. International Conference on Image Processing*, volume 1, pages 449–452, 2002.

[81] Tony Lindeberg. Feature detection with automatic scale selection. *International journal of computer vision*, 30(2):79–116, 1998.

[82] Michael Livingston. Alcohol outlet density and assault: a spatial analysis. *Addiction*, 103(4):619–628, 2008.

[83] Michael Livingston. Alcohol outlet density and harm: comparing the impacts on violence and chronic harms. *Drug and alcohol review*, 30(5):515–523, 2011.

[84] K. Lloyd, P. L. Rosin, A. D. Marshall, and S. C. Moore. Violent behaviour detection using local trajectory response. In *7th International Conference on Imaging for Crime Detection and Prevention (ICDP 2016)*, pages 1–6, Nov 2016.

[85] Kaelon Lloyd, Paul L. Rosin, David Marshall, and Simon C. Moore. Detecting violent and abnormal crowd activity using temporal analysis of grey level co-occurrence matrix (GLCM)-based texture measures. *Machine Vision and Applications*, 28(3-4):361–371, 2017.

[86] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157, 1999.

[87] Bruce D. Lucas and Takeo Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, pages 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc.

[88] Penousal Machado and Amílcar Cardoso. Computing aesthetics. *Advances in artificial intelligence*, pages 105–119, 1998.

[89] Penousal Machado, Juan Romero, Marcos Nadal, Antonino Santos, João Correia, and Adrián Carballal. Computerized measures of visual complexity. *Acta psychologica*, 160:43–57, 2015.

[90] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981, 2010.

[91] A. Marana, L. da Costa, R. Lotufo, and S. Velastin. On the Efficacy of Texture Analysis for Crowd Monitoring. In *Proceedings of the International Symposium on Computer Graphics, Image Processing, and Vision*, SIBGRAPHI, pages 354–361, Washington, DC, USA, 1998. IEEE Computer Society.

[92] Jorge S. Marques, Pedro M. Jorge, Arnaldo J. Abrantes, and J. M. Lemos. Tracking Groups of Pedestrians in Video Sequences. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, volume 9, pages 101–101, 2003.

[93] Mark Marsden, Kevin McGuinness, Suzanne Little, and Noel E. O'Connor. Holistic features for real-time crowd behaviour anomaly detection. *IEEE International Conference on Image Processing*, pages 918–922, 2016.

[94] Mark Marsden, Kevin McGuinness, Suzanne Little, and Noel E. O'Connor. ResnetCrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification. In *Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on*, pages 1–7. IEEE, 2017.

[95] Clara Martimportugués-Goyenechea and Luis Gómez-Jacinto. Simultaneous multiple stressors in the environment: Physiological stress reactions, performance, and stress evaluation. *Psychological reports*, 97(3):867–874, 2005.

[96] Krešimir Matković, László Neumann, Attila Neumann, Thomas Psik, and Werner Purgathofer. Global Contrast Factor - A New Approach to Image Contrast. In *Proceedings of the First Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging*, Computational Aesthetics'05, pages 159–167, Aire-la-Ville, Switzerland, Switzerland, 2005. Eurographics Association.

[97] A. Mecocci and F. Micheli. Real-time recognition of violent acts in monocular colour video sequences. In *2007 IEEE Workshop on Signal Processing Applications for Public Security and Forensics*, pages 1–4, April 2007.

[98] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behaviour detection using social force model. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942, 2009.

[99] Zihan Meng, Jiabin Yuan, and Zhen Li. Trajectory-pooled deep convolutional networks for violence detection in videos. In *International Conference on Computer Vision Systems*, pages 437–447. Springer, 2017.

[100] James C. Miller, Matthew L. Smith, and Michael E. McCauley. Crew fatigue and performance on US coast guard cutters. Technical report, U.S. Coast Guard Research and Development Center, 1998.

[101] Simon C. Moore, Mario Flajšlik, Paul L. Rosin, and David Marshall. A particle model of crowd behavior: Exploring the relationship between alcohol, crowd dynamics and violence. *Aggression and Violent Behavior*, 13(6):413–422, 2008.

[102] Yunyoung Nam and Sangjin Hong. Real-time abnormal situation detection based on particle advection in crowded scenes. *Journal of Real-Time Image Processing*, 10(4):771–784, Dec 2015.

[103] Randy J. Nelson and Brian C. Trainor. Neural mechanisms of aggression. *Nature Reviews Neuroscience*, 8(7):536–546, 2007.

[104] Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno Garcia, and Rahul Sukthankar. Violence Detection in Video Using Computer Vision Techniques. *International conference on Computer analysis of images and patterns*, pages 332–339, 2011.

[105] Clive Norris and Gary Armstrong. *The Maximum Surveillance Society: The Rise of CCTV*, volume 49. Oxford, 1999.

[106] Clive Norris, Mike Mccahill, and David Wood. Editorial . The Growth of CCTV : a global perspective on the international diffusion of video surveillance in publicly accessible space . *Surveillance & Society*, 2(2/3):110–135, 2004.

[107] Eric Nowak, Frédéric Jurie, and Bill Triggs. Sampling strategies for bag-of-features image classification. *Computer Vision–ECCV 2006*, pages 490–503, 2006.

[108] Srinivas Palamarthy, Hani S. Mahmassani, and Randy B. Machemehl. Models of pedestrian crossing behavior at signalized intersections. Technical report, University of Texas at Austin. Center for Transportation Research, 1994.

[109] Raja Parasuraman, Ewart de Visser, Ellen Clarke, W. Ryan McGarry, Elizabeth Hussey, Tyler Shaw, and James C. Thompson. Detecting threat-related intentional actions of others: effects of image quality, response mode, and target cuing on vigilance. *Journal of experimental psychology: applied*, 15(4):275, 2009.

[110] Nick Perham, Simon C. Moore, Jonathan Shepherd, and Bryany Cusens. Identifying drunkenness in the night-time economy. *Addiction*, 102(3):377–380, 2007.

[111] Enrico Quagliarini, Gabriele Bernardini, Luca Spalazzi, et al. EPES–earthquake pedestrians' evacuation simulator: A tool for predicting earthquake pedestrians' evacuation in urban outdoor scenarios. *International journal of disaster risk reduction*, 10:153–177, 2014.

[112] Aravinda S. Rao, Jayavardhana Gubbi, Sutharshan Rajasegarar, Slaven Marusic, and Marimuthu Palaniswami. Detection of anomalous crowd behaviour using hyperspherical clustering. *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8, 2014.

[113] Pedro Canotilho Ribeiro, Romaric Audigier, and Quoc Cuong Pham. RIMOC, a feature to discriminate unstructured motions: Application to violence detection for video-surveillance. *Computer Vision and Image Understanding*, 144:121–143, 2016.

[114] Ira J. Roseman. Appraisals, rather than unpleasantness or muscle movements, are the primary determinants of specific emotions. *Emotion*, 2004.

[115] David Ryan, Simon Denman, Clinton Fookes, and Sridha Sridharan. Textures of optical flow for real-time anomaly detection in crowds. In *2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 230–235, Aug 2011.

[116] David Salomon. *Data compression: the complete reference*. Springer Science & Business Media, 2004.

[117] Christopher Sarno. The impact of closed circuit television on crime in Sutton town centre. *Towards a safer Sutton*, pages 13–49, 1996.

[118] Nima Sarshar, Mahmoud R. Halfawy, and Jantira Hengmeechai. Video processing techniques for assisted CCTV inspection and condition rating of sewers. *Journal of Water Management Modeling*, pages 129–147, 2009.

[119] Nick Scott, Michael Livingston, Aaron Hart, James Wilson, David Moore, and Paul Dietze. SimDrink: an agent-based NetLogo model of young, heavy drinkers for conducting alcohol policy experiments. *Journal of Artificial Societies and Social Simulation*, 19(1):10, 2016.

[120] Hamid R. Sheikh and Alan C. Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006.

[121] Hamid R. Sheikh, Alan C. Bovik, and Gustavo De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on image processing*, 14(12):2117–2128, 2005.

[122] Lawrence W. Sherman and Dennis P. Rogan. Effects of gun seizures on gun violence: “Hot spots” patrol in Kansas City. *Justice Quarterly*, 12(4):673–693, 1995.

[123] Lawrence W. Sherman and David Weisburd. General deterrent effects of police patrol in crime “hot spots”: A randomized, controlled trial. *Justice quarterly*, 12(4):625–648, 1995.

[124] Feng Shi, Emil Petriu, and Robert Laganiere. Sampling strategies for real-time action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2595–2602, 2013.

[125] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 746–760, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[126] Bernard W. Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

[127] V. Sivarajasingam, J. P. Shepherd, and K. Matthews. Effect of urban closed circuit television on assault injury and violence detection. *Injury prevention*, 9(4):312–316, 2003.

[128] Eric Sundstrom. An experimental study of crowding: effects of room size, intrusion, and goal blocking on nonverbal behavior, self-disclosure, and self-reported stress. *Journal of Personality and Social Psychology*, 32(4):645, 1975.

[129] Henry A. Swett, Paul R. Fisher, Aaron I. Cohn, Perry L. Miller, and Pradeep G. Mutalik. Expert system-controlled image display. *Radiology*, 172(2):487–493, 1989.

[130] Henry A. Swett and Perry L. Miller. ICON: a computer-based approach to differential diagnosis in radiology. *Radiology*, 163(2):555–558, 1987.

[131] Ralph B/ Taylor and Stephen Gottfredson. Environmental design, crime, and prevention: An examination of community dynamics. *Crime and justice*, 8:387–416, 1986.

[132] Mercè Teixidó, Tomàs Pallejà, Marcel Tresanchez, Miquel Nogués, and Jordi Palacín. Measuring oscillating walking paths with a LIDAR. *Sensors*, 11(5):5071–5086, 2011.

[133] A. H. Tickner and E. C. Poulton. Monitoring up to 16 synthetic television pictures showing a great deal of movement. *Ergonomics*, 16(4):381–401, 1973.

[134] Carlo Tomasi and Takeo Kanade. Detection and Tracking of Point Features. Technical report, International Journal of Computer Vision, 1991.

[135] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1521–1528. IEEE, 2011.

[136] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[137] Ioannis Tziakos, Andrea Cavallaro, and Li-Qun Xu. Local abnormality detection in video using subspace learning. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pages 519–525. IEEE, 2010.

[138] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. scikit-image: image processing in Python. *PeerJ*, 2:e453, 6 2014.

[139] G. van Voorthuijsen, H. van Hoof, M. Klima, K. Roubik, M. Bernas, and P. Pata. CCTV effectiveness study. *Proceedings 39th Annual 2005 International Carnahan Conference on Security Technology*, 2005.

[140] Paul Virilio. *The vision machine.* Indiana University Press, 1994.

[141] Blake Byron Walker, Nadine Schuurman, and S Morad Hameed. A gis-based spatiotemporal analysis of violent trauma hotspots in vancouver, canada: identification, contextualisation and intervention. *BMJ open*, 4(2):e003642, 2014.

[142] Bobo Wang, Hong Bao, Shan Yang, and Haitao Lou. Crowd density estimation based on texture feature extraction. *Journal of Multimedia*, 8(4):331–337, aug 2013.

[143] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3551–3558. IEEE, 2013.

[144] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC 2009-British Machine Vision Conference*, pages 124–1. BMVA Press, 2009.

[145] Jing Wang and Zhijie Xu. Spatio-temporal texture modelling for real-time crowd anomaly detection. *Computer Vision and Image Understanding*, 144:177–187, 2016.

[146] Tian Wang and Hichem Snoussi. Detection of abnormal events via optical flow feature analysis. *Sensors*, 15(4):7156–7171, mar 2015.

[147] Tianjiao Wang, Jianping Wu, Pengjun Zheng, and Mike McDonald. Study of pedestrians' gap acceptance behavior when they jaywalk outside crossing facilities. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 1295–1300. IEEE, 2010.

[148] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[149] Alison L. Warburton and Jonathan P. Shepherd. Tackling alcohol related violence in city centres: effect of emergency medicine and police intervention. *Emergency Medicine Journal*, 23(1):12–17, 2006.

[150] Brandon C. Welsh and David P. Farrington. *Crime prevention effects of closed circuit television: a systematic review*, volume 252. Citeseer, 2002.

[151] Brandon C. Welsh and David P. Farrington. Effects of closed-circuit television on crime. *The Annals of the American Academy of Political and Social Science*, 587(1):110–135, 2003.

[152] Shandong Wu, Brian E. Moore, and Mubarak Shah. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2054–2060, 2010.

[153] Mei-Ling Xiao, Yang Chen, Ming-Jiao Yan, Liao-Yuan Ye, and Ben-Yu Liu. Simulation of household evacuation in the 2014 ludian earthquake. *Bulletin of Earthquake Engineering*, 14(6):1757–1769, 2016.

[154] Long Xu, Chen Gong, Jie Yang, Qiang Wu, and Lixiu Yao. Violent video detection based on MoSIFT feature and sparse coding. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 3538–3542, 2014.

[155] Hanxuan Yang, Ling Shao, Feng Zheng, Liang Wang, and Zhan Song. Recent advances and trends in visual tracking: A review. *Neurocomputing*, 74(18):3823–3831, 2011.

[156] Wenjian Yu and Anders Johansson. Modeling crowd turbulence by many-particle simulations. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 76(4):1–11, 2007.

[157] Fabian Zambrano, Pablo Concha, Francisco Ramis, Liliana Neriz, Maria Bull, Patricio Veloz, and Jaime Carvajal. Improving patient access to a public hospital complex using agent simulation. In *Winter Simulation Conference (WSC), 2016*, pages 1277–1288. IEEE, 2016.

[158] Limao Zhang, Mengjie Liu, Xianguo Wu, and Simaan M. AbouRizk. Simulation-based route planning for pedestrian evacuation in metro stations: A case study. *Automation in Construction*, 71:430–442, 2016.

[159] Xiaoping Zheng, Tingkuan Zhong, and Mengting Liu. Modeling crowd evacuation of a building based on seven methodological approaches. *Building and Environment*, 44(3):437–445, 2009.

[160] Shifu Zhou, Wei Shen, Dan Zeng, and Zhijiang Zhang. Unusual event detection in crowded scenes by trajectory analysis. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 1300–1304, 2015.

[161] Liqi Zhu, Dennis M. Gorman, and Scott Horel. Alcohol outlet density and violence: a geospatial analysis. *Alcohol and alcoholism*, 39(4):369–375, 2004.

[162] Xiaobin Zhu, Jing Liu, Jinqiao Wang, Wei Fu, and Hanqing Lu. Weighted interaction force estimation for abnormality detection in crowd scenes. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7726 LNCS(PART 3):507–518, 2013.

[163] Guang Yong Zou. Toward using confidence intervals to compare correlations. *Psychological methods*, 12(4):399, 2007.

[164] Karel Zuiderveld. Contrast limited adaptive histogram equalization. *Graphics gems*, pages 474–485, 1994.