Target Classification in Multimodal Video

Iain Rodger

A thesis submitted in fulfillment of the degree of

Engineering Doctorate

Engineering & Physical Sciences
Institute of Sensors, Signals & Systems

Heriot Watt University

Land and Air Systems

Thales UK
Glasgow

October 2017

Abstract

The presented thesis focuses on enhancing scene segmentation and target recognition methodologies via the mobilisation of contextual information. The algorithms developed to achieve this goal utilise multi-modal sensor information collected across varying scenarios, from controlled indoor sequences to challenging rural locations. Sensors are chiefly colour band and long wave infrared (LWIR), enabling persistent surveillance capabilities across all environments. In the drive to develop effectual algorithms towards the outlined goals, key obstacles are identified and examined: the recovery of background scene structure from foreground object 'clutter', employing contextual foreground knowledge to circumvent training a classifier when labeled data is not readily available, creating a labeled LWIR dataset to train a convolutional neural network (CNN) based object classifier and the viability of spatial context to address long range target classification when *big data* solutions are not enough.

For an environment displaying frequent foreground clutter, such as a busy train station, we propose an algorithm exploiting foreground object presence to segment underlying scene structure that is not often visible. If such a location is outdoors and surveyed by an infra-red (IR) and visible band camera set-up, scene context and contextual knowledge transfer allows reasonable class predictions for thermal signatures within the scene to be determined.

Furthermore, a labeled LWIR image corpus is created to train an infrared object classifier, using a CNN approach. The trained network demonstrates effective classification accuracy of 95% over 6 object classes. However, performance is not sustainable for IR targets acquired at long range due to low signal quality and classification accuracy drops. This is addressed by mobilising spatial context to affect network class scores, restoring robust classification capability.

Acknowledgements

Writing this tome, my magnum opus if you will, would not have been possible without the support of a great number of people over the last 5 years and beyond.

Firstly, I am eternally grateful to my family for their encouragement throughout my education. Specifically, I am indebted to my mother for being brought up in an environment that valued learning and knowledge. This led to my natural inquisitiveness which is invaluable while going through a doctorate. Plus, if ever I needed to have a mad weekend there really is no place like home!

At Thales, Adam Sroka is most deserving of my thanks for being not only a sponge for my moaning, but also a soundboard for ideas. I'm so glad we were in the same boat together, misery loves company. What an *utter* lad.

Special mentions have to go to my supervisors, Barry Connor and Neil Robertson, for providing just the right input when I most needed it. We got there eventually.

To my beloved GSC - Fin, Mick & Arran, thanks for listening to my constant whinging throughout the entire process and just generally being about when I needed an escape from reality. Also, many thanks for helping out with my wedding as I'm not the most organised person, especially in the endgame of a doctorate. It shall not be forgotten!

I am also grateful to my work colleagues at Bridgeall Libraries for the brief time I was a resident, if it wasn't such a great place to work I doubt I would have had the energy to put the thesis to bed in my free time. Iain & Dan, you are great colleagues, teachers and friends. Long may it continue. Thanks for giving me a shot and respecting what I do, it was some ride. EAT THE CHEESE!!!

Lastly, thanks to my darling wife. Words cannot capture the support you have given me throughout the EngD. Achieving this would not have been possible otherwise. We did it!!

Publications

- *Recovering Background Regions in Videos of Cluttered Urban Scenes*, I. Rodger, B. Connor & N.M. Robertson, in IEEE International Conference on Image Processing (ICIP), 2015
- *Multi-modal Object Detection via Contextual Foreground Regions Background Regions* , I. Rodger, B. Connor & N.M. Robertson, in Proceedings of 4th IMA, Mathematics in Defence Conference, 2015
- *Classifying objects in LWIR imagery via CNNs*, I. Rodger, R. Abbott, B. Connor & N.M. Robertson, in Proceedings of SPIE Security & Defence, 9987, Electro-Optical and Infrared Systems; Technology and Applications, 2016
- *Enhancing long range ATR using spatial context*, I. Rodger, B. Connor & N.M. Robertson, in Sensor Signal Processing for Defence (SSPD), 2017 [$Tobepresented$]

Awards

- Early Career Award for presenting *Multi-modal Object Detection via Contextual Foreground Regions Background Regions* at the Institute of Mathematics & its Applications (IMA) Conference on Mathematics in Defence, held in Oxford 2015.
- Best Student paper presentation at the conference on Electro-Optical and Infrared Systems; Technology and Applications, part of the SPIE Security and Defence International Symposia 2016, held in Edinburgh in September 2016

HERIOT WATT UNIVERSITY

ACADEMIC REGISTRY
Research Thesis Submission

Name*:*	Iain Rodger		
School/PGI:	Institute of Sensors, Signals and Systems		
Version: *(i.e. First, Resubmission, Final)*	Final	Degree Sought (Award **and** Subject area)	Doctor of Engineering

Declaration

In accordance with the appropriate regulations I hereby submit my thesis and I declare that:

1) the thesis embodies the results of my own work and has been composed by myself
2) where appropriate, I have made acknowledgement of the work of others and have made reference to work carried out in collaboration with other persons
3) the thesis is the correct version of the thesis for submission and is the same version as any electronic versions submitted*.
4) my thesis for the award referred to, deposited in the Heriot-Watt University Library, should be made available for loan or photocopying and be available via the Institutional Repository, subject to such conditions as the Librarian may require
5) I understand that as a student of the University I am required to abide by the Regulations of the University and to conform to its discipline.

* *Please note that it is the responsibility of the candidate to ensure that the correct version of the thesis is submitted.*

Signature of Candidate*:*		Date:	

Submission

Submitted By *(name in capitals):*	
Signature of Individual Submitting:	
Date Submitted:	

For Completion in the Student Service Centre (SSC)

Received in the SSC by *(name in capitals):*			
Method of Submission *(Handed in to SSC; posted through internal/external mail):*			
*E-thesis Submitted (**mandatory for final theses)***			
Signature:		Date:	

Please note this form should be bound into the submitted thesis.

Updated February 2008, November 2008, February 2009, January 2011

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

BTG	Bush Tree Grass
AFGBG	Aggregate Foreground-Background
AI	Artificial Intelligence
ANN	Artificial Neural Network
ATDR	Automatic Target Detection & Recognition
ATD	Automatic Target Detection
ATR	Automatic Target Recognition
BGFG	Background-Foreground
BG	Background
CCTV	Closed Circuit Television
CIE	International Commission on Illumination
CNN	Convolutional Neural Network
CRF	Conditional Random Fields
CR	Cover Rate
DAGS	Stanford Background Dataset
DMZ	De-Militarised Zone
EGB	Efficient Graph-Based
EM	Electromagnetic

EngD	Engineering Doctorate
EPSRC	Engineering and Physical Sciences Research Council
FGBG	Foreground-Background
FG	Foreground
FIR	Far Infrared
HOG	Histogram of Oriented Gradients
IR	Infrared
ISSS	Institute for Sensors, Signals and Systems
JaccD	Jaccard Distance
LDA	Linear Discriminant Analysis
LiDAR	Light Detection And Ranging
LWIR	Longwave Infrared
MP	Megapixel
MSER	Maximally Stable Extremal Regions
MST	Minimum Spanning Tree
MWIR	Mid-wave Infrared
NCut	Normalised Cuts
NIR	Near Infrared
NMS	Non-Maximum Suppression
NN	Neural Network
QWIP	Quantum Well Infrared Photodetector
R&D	Research and Development
ReLU	Rectified Linear Unit
RGB	Red-Green-Blue Colour Model
RNN	Recurrent Neural Networks

ROI	Region of Interest
SLIC	Simple Linear Iterative Clustering
SNR	Signal-to-Noise Ratio
SP	Superpixel
SVM	Support Vector Machines
SWIR	Short-wave Infrared
TBG	Total Background Layer
TI	Thermal Imager
UAV	Unmanned Aerial Vehicle
VOI	Variation of Information

Chapter 1
Introduction

Humans use the eye as a sensor to collect visual data from which our brain extracts useful information. This allows us to effectively understand the surrounding world environment whilst generating a conscious perceptual experience. These experiences form part of our visual system and help the brain understand future scenes, meaning humans are very efficient at quickly comprehending complex visual events with relative ease. The ultimate goal of computer vision is to develop methods that allow computers or machines to intelligently sense and understand surroundings akin to the human visual system, learning from prior information and experiences in the process. If this goal is fully realised the potential applications are vast across a wide range of industries. From the impending wave of autonomous driverless vehicles to robotic care assistants for the elderly, advanced and reliable computer vision systems will be integral to such systems.

The security and defence industry is becoming increasingly reliant on intelligent signal processing methods to achieve *persistent surveillance* capabilities, where systems must be able to perform 24-hours a day. Often the aim is to maximise situational awareness in military or surveillance based tasks. Put simply, situational awareness relates to understanding critical information about surrounding events and how it can affect future decision making in dynamic scenarios [19]. Given the limitations of human processing capabilities, employing human operators on massive-scale surveillance systems is infeasible. By researching and developing tools to flag potential threats in a scene, the burden placed on humans in normal to dangerous environments can be alleviated. Computer vision processes are utilised to achieve such improvements providing solutions that a person could simply not replicate in terms of execution speed or performance, regardless of operator skill or training. An illustrative example of this concept is provided in Figure 1.1.

In Figure 1.1 a typical urban night scene is presented, captured using a visible

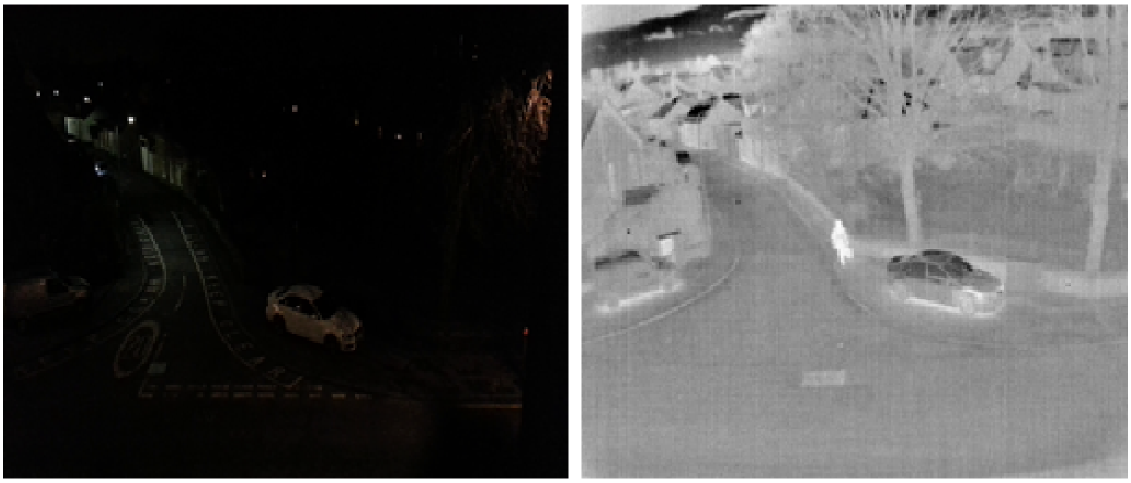


Figure 1.1: The left hand image shows a colour image capturing an urban street scene in Glasgow. Given the poor illumination and sensor band it is hard to discern scene content. The right hand image captures the same view but using a low quality thermal imager, revealing a person in the centre of the scene.

and thermal camera. Regardless of how many analysts observe the colour image, it cannot possibly be ascertained if there is a credible threat. When the extra thermal band information is presented, however, only then is the presence of a pedestrian revealed. Exploiting the persistent surveillance capability offered by thermal sensors is a key focus of this research project. The real world consequence from deploying robust signal processing systems, in a typical defence environment, is an enhanced situational awareness leading to more informed decision making overall.

In terms of signal acquisition the human eye is limited to sensing light only in the visible band of the Electromagnetic (EM) spectrum. Light waves enter the eye activating specialised rod and cone cells within the retinal structure, creating electrical impulses transmitted to the brain for comprehension. This general process is very similar to the collection of light information using engineered systems, where EM signals propagate through an optical system until they reach a *detector*. Depending on the material used in the detector slice the absorbed light will generate electrical signals to be interpreted, for instance in the form of images. Thus, a signal processing strategy to aid scene understanding is required. Developing such methods is the focus of this research thesis.

Robust and intelligent image comprehension is hard to achieve using machines. By comparison it is a simple task for humans to perform, where our visual system benefits from millions of years of evolution. The human visual system is remarkably good at processing light information, quickly forming scene understanding and building inferences in real-time. This adaptable learning mechanism is powerful, especially when faced with new, effectively *unseen* information. Conversely, such a robust signal processing system, capable of transferable learning, is still not fully realised using human engineered hardware and software. Through effective sensor utilisation it is possible to build systems capable of 24-hour operation and sense more than one waveband, such as thermal infrared (IR) , addressing the fundamental limitations of the human visual system. This thesis is concerned with replicating the transferable inference power of the brain towards computer algorithms, deployable with multimodal sensor suites. Contextual models and recent machine learning tools are employed, providing the key mechanism to transfer implicit scene understanding to our algorithms.

1.1 Engineering Doctorate Program

The presented thesis examines and discusses research conducted towards attaining an Engineering Doctorate (EngD) in Applied Photonics. The undertaken work operates at the very cusp of this research remit, in the field of computer vision. The central focus is on the development of image processing techniques to enhance target classification and scene understanding, via the integration of contextual information. The project was carried out primarily within a sponsor company, Thales UK, and the Visionlab, part of the Institute for Sensors, Signals and Systems (ISSS) based at Heriot-Watt University, Edinburgh. Project funding was provided jointly by the Engineering and Physical Sciences Research Council (EPSRC) and an industrial partner, Thales UK. The EngD is centrally managed by the Centre for Doctoral Training in Applied Photonics, formally the Industrial Doctorate Centre for Optics and Photonics Technologies, via a partnership of 100 academic research groups spread across 5 universities. A core advantage of such a scheme is the generated research has a high relevance towards industrial applications. This EngD project exemplifies the industrial relevance benefit, where Thales UK provide resources in exchange for novel solutions towards building robust detection and classification algorithms, developed using their state-of-the-art thermal imagers. The industrial motivation is expanded further in Section 1.2.

This introductory chapter captures the key research themes present throughout the thesis and is laid out as follows: Section 1.2 outlines the overall problem statement and underlying motivation for conducting the research. This is considered from not only the academic viewpoint, but also from a current industrial position to relay the commercial value of the research. Section 1.3 presents the specific aims of the EngD project while Section 1.4 follows with the transfer of knowledge to the defence industry. Lastly, our research contributions are defined in Section 1.5 and the remaining thesis is outlined in Section 1.6.

1.2 Motivation

By examining the stimulating factors underpinning the research, the scope of the problem domain and surrounding context can be grasped. This can then be related to current commercial interests, specifically in security and defence. Lets begin by initially providing a hypothetical scenario to capture the true essence of the problem, from an academic aspect. The motivational scene relates to problems faced today in defence and surveillance operations, which industrial bodies are keen to address for reasons later discussed. Moreover, the research driving force can be elaborated upon from this industrial perspective and challenges faced by Thales UK can be focused on in doing so. It is important to recall that both of these viewpoints must be considered when evaluating an EngD. Academically the research must be novel and relevant, but it also needs to address current gaps in industry where potential solutions are desirable. In other words the project deliverables are likely to be applicable to industry relevant problems, leading to new products that ultimately would not be possible without such commercially focused research. This is a holistic view of the EngD. Let us now discuss the academic problem statement and wider motivations.

1.2.1 Academic & Industrial Motivation

For Thales, their customers and the wider research community, the task to overcome is concerned with scene segmentation and target recognition methodologies, specifically in a surveillance based context. At the outset of this project such systems were demonstrated to perform poorly in complex and dynamic scenes, where changing illumination and occlusions are frequent. While great strides have been made to tackle this problem over recent years, there remains a gap for reliable and robust scene perception and target recognition approaches for 24-hour surveillance. To grasp the importance of such a capability *situational awareness* and how these methods relate must be considered.

Situational awareness is the foremost concern when human decisions have to be made with potentially significant consequences in a specific problem domain [19]. A relatable example can be described by contemplating a tourist visiting a foreign country. In this case a tourist, behaving in a normal manner, shall strive to remain out of harms way and avoid risky encounters. Due to the unfamiliar nature of foreign surroundings, humans will instinctively be more alert and perceptive to environmental factors that may compromise such a goal. It is only when a person becomes at ease with the local setting that vigilance dissipates. This enhanced interpretation of the environment in proximity and how it may affect future events, as well as decisions, is the essence of situational awareness. At the local level described in such a scenario, the surrounding environment must be observed in some fashion to actually garner information to aid the decision making process.

If the the idea of situational awareness is extended to the realm of military scenarios, where local environments may be hazardous and threatening, or potential threats may exist at long ranges and actively try to avoid detection, obtaining enough useful knowledge about surroundings can be fraught with such challenges leading to impaired threat perception. Moreover, monitoring equipment tends to be *passive* where sensors only collect rather than emit physical signals, to detect nearby objects for instance. Such sensors must be passive in a hostile environment, typical of military scenarios, as an active sensor emission can be detected by enemy forces which is not ideal for covert surveillance units. This is the chief benefit of employing passive sensors in such circumstances, although it is a tradeoff between being undetectable and gaining enhanced signals from active sensing technology. Utilising these systems is critical for gathering information to improve local situational awareness and the capacity to make informed decisions. This is why current and next generation land vehicles of the British Army are outfitted with such sensor systems, as shown in Figure 1.2.

The work presented in this thesis is only concerned with image information captured by sensors measuring light from visible and infrared portions of the EM spectrum. The visual band is what humans naturally perceive and is an obvious choice for this reason. However, colour cameras are limited by dynamic illumination and cannot effectively see in the absence of adequate light sources. Infrared band cameras address this limitation and can sense any hot bodies emitting thermal radiation, making them illumination invariant. In other words IR sensors can effectively *see in the dark*. Only longwave infrared (LWIR) is considered as the thermal band of interest for any research conducted, since it is emission dominated compared to other portions of the IR band.

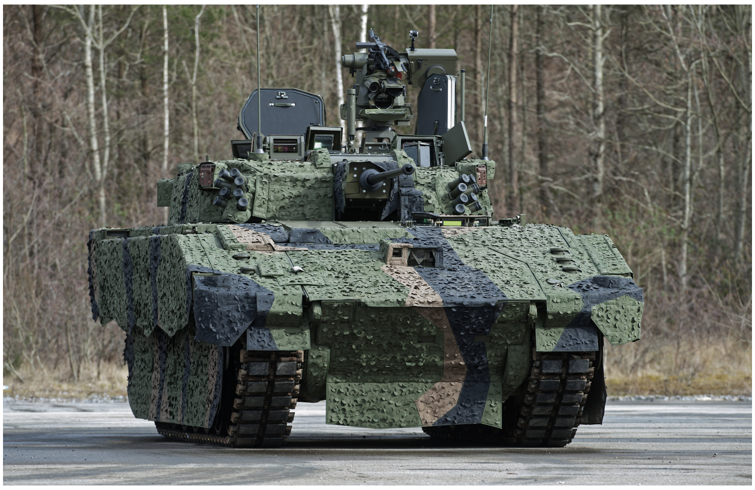


Figure 1.2: The figure presents a prototype Ajax armoured fighting vehicle, soon to be commissioned for use by the British Army. This new breed of vehicle is equipped with state-of-the-art sensors operating in multiple wavebands, as well as local situational awareness systems. One of the key capabilities of the new vehicle design is the electronic architecture and on-board data capture system, as intelligent processing and information analysis is critical to obtain enhanced situation awareness. Image re-used courtesy of ©Crown Copyright.

Efficient night-vision greatly improves situational awareness and offers a powerful capability for persistent surveillance. This advantage was demonstrated in the first Gulf War, where land vehicles designed towards the end of the Cold War were employed to great effect. These combat vehicles were fitted with early stage thermal imagers (TI) enabling British and American forces to fight under the cover of darkness, where opposing forces could not operate as they lacked this key technology. This asymmetric war signposted how critical persistent surveillance capabilities can be, where enhanced situational awareness provides a distinct advantage. Thales UK have considerable expertise manufacturing high quality TIs and one of their key products is utilised throughout the thesis, the Catherine MP [20].

The Catherine MP is a state-of-the-art TI operating in the LWIR band and is the primary choice TI of the British Army. It can capture very detailed imagery across a large range of distances, making it an ideal selection for our research purposes as well as maintaining an obvious link to industry. The camera itself and example imagery are shown in Figure 1.3. However, the caveat to 24-hour sensing technology in a domain where vigilance is required has not yet been discussed.

Let us consider a plausible scenario where a vehicle, similar to one as shown in Figure 1.2, is likely to operate. A machine equipped with an array of sensors and

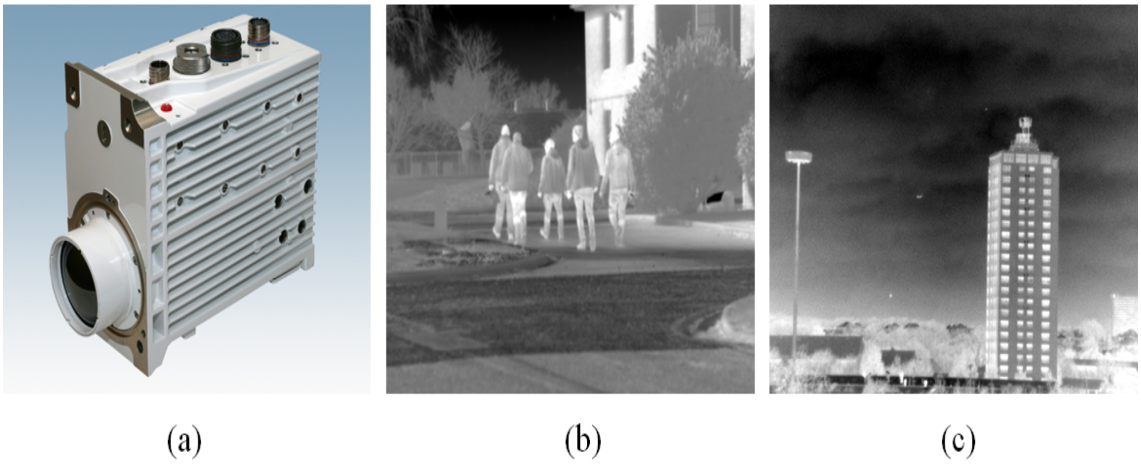


Figure 1.3: Image (a) shows the Catherine MP LWIR TI. Image (b) captures a crowd of pedestrians at close range while (c) is a wide-angle shot of an urban vista, illustrating an aircraft passing across the sky in the very far distance.

intelligent processing is designed for performing surveillance of targets, from close to far ranges, while stationery and on the move. It will most probably undertake such tasks deployed in a hostile environment under sustained threat, requiring constant alertness from operators within the vehicle. This is especially true in a modern battle domain where thermal sensing technology is now commonplace, allowing all armed forces 24-hour surveillance capabilities.

The load placed on human operators in such high pressure conditions is immense. Performance on tasks requiring long periods of focus, such as reviewing several surveillance feeds from multiple cameras, is known to deteriorate over time. Part of this is due to the limited capacity of the human visual system for processing information, which is further compounded by the often long and repetitive nature of surveillance type work [21, 22]. This performance degradation can result in poor decisions even with advanced situational awareness, leading to potentially disastrous consequences. Thus, it would be highly desirable to relieve system operators from such a burden by replacing or augmenting some of their human visual process using intelligent signal processing. Ultimately, robust algorithms for improving automatic detection and recognition tasks are highly desired, where any successful approaches will alleviate the burden from human operators. The problem is not confined to the security and defence domain, but is also relevant for civilian applications where autonomous scene perception is required.

1.3 Aims

Having considered the academic and industrial drivers for this project, within the context of a hypothetical surveillance situation, the goals of the EngD as a whole can now be stated. Ideal solutions presented shall incorporate LWIR sensor information and be applicable from static platforms. Operating under these overarching criteria, this work aims to provide answers to the following questions:

1. Can the presence of foreground objects influence scene perception and provide a route to persistent surveillance?
2. Does the mobilisation of scene specific context, gained via semantic segmentation, enhance target recognition performance using multimodal sensor systems?

The first question is explored and answered in Chapters 3 and 4 respectively, where the the presence of foreground objects relates to the underlying scene structure. This information is exploited by a knowledge transfer approach, allowing objects in thermal imagery to be classified by reference observations in corresponding colour band surveillance imagery. The second question is considered in Chapters 5 and 6 respectively, where a robust target classifier for LWIR images using recent machine learning techniques is developed. This classifier is then applied to a challenging defence relevant problem, showing performance is considerably improved by augmenting classification scores with contextual knowledge.

1.3.1 Scope & Assumptions

To present the body of work fairly it is reasonable to outline any key assumptions and limitations made throughout, with regards to the stated aims. As such, any methods presented are applicable only to static camera surveillance scenarios. In these situations, the main goal will be to determine what is contained within the scene, whether it be object classes or the region types. Furthermore, the thesis focuses on colour and thermal band sensor information. Techniques to extract underlying scene structure or segment semantic regions are developed only on colour imagery. Lastly, only LWIR imagery is incorporated with any proposed methods utilising thermal band information.

1.4 Knowledge Transfer

1.4.1 Research Outputs

- A conference paper [23] was presented at the IEEE International Conference on Image Processing, 2015. This work is explored in Chapter 3.
- An award winning paper presentation was delivered at the Institute of Mathematics & its Applications 4th Conference on Mathematics in Defence, held in Oxford 2015. This is discussed further in Chapter 4.
- This was followed by another award winning conference paper [16] and presentation at SPIE Security & Defence, 2016.
- Lastly, another paper and presentation [24] is due to be given in 2017 at SSPD in London. The contents of both SPIE and SSPD publications are covered in Chapters 5 and 6.

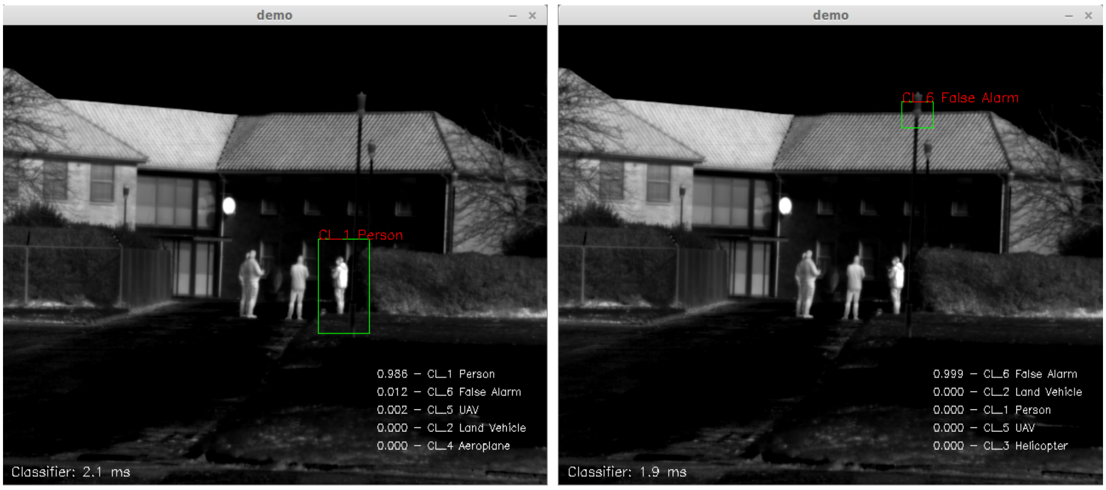
1.4.2 Commercial Impact

Key strands of research from this project have been presented many times internally to senior members of the Thales UK technology directorate, who operate at a strategic level guiding future areas of research and development (R&D) . In 2016, recent work on LWIR target detection was presented at a Thales UK Innovation Day. This is a high profile internal event where various ranks of personnel observe current projects within Thales.

The main results of the LWIR classification scheme from [16] were shown alongside a real-time demo. Example illustrations of this highly accurate object recognition algorithm are shown in Figure 1.4. This demo and *exhibition stall* won Best Innovation Stand at the event, but more importantly the outcomes of this thesis have had a significant impact on the future research direction within Thales Land and Air Systems. For such a large company this is a remarkable achievement for a single EngD student to accomplish.

1.5 Contributions

Overall, the work presented in this thesis serves as a useful indicator for the rise of machine learning within computer vision. While it always had a significant place in



```
0.986 – CL_1 Person
0.012 – CL_6 False Alarm
0.002 – CL_5 UAV
0.000 – CL_2 Land Vehicle
0.000 – CL_4 Aeroplane
Classifier: 2.1 ms
```

(a)

```
0.999 – CL_6 False Alarm
0.000 – CL_2 Land Vehicle
0.000 – CL_1 Person
0.000 – CL_5 UAV
0.000 – CL_3 Helicopter
Classifier: 1.9 ms
```

(b)

Figure 1.4: This pair of LWIR thermal images from the Catherine MP are fed into a trained object classifier. The user is allowed to draw a region of interest around the target, shown as a green box in both cases. Output classifier probabilities for prediction confidence are shown in the bottom right hand corner. The selected object classified in image (a) is correctly recognised as a person. Image (b) illustrates the correct classification of a false alarm, reflected in the softmax output in the bottom right hand corner of the image.

the field, most research made incremental gains employing such methods. However, with the advent of recent deep learning approaches the field has exploded in terms of performance and interest. Machine learning became hugely popular for segmentation and classification tasks, under the supervised learning umbrella from 2014 onwards, a mere two years or less after the seminal early works reviving interest in this area. Overall the step-change in thought and approach towards computer vision problems is reflected in the layout, with early data chapters trying to gain object classification for *free* via scene context and later chapters fully adopting deep neural network methods. In other words, it is a relevant snapshot of the field over time.

The essential contributions outlined in this thesis can be summarised in the following ways:

- Developing a scene segmentation process influenced by pedestrians and the presence of other foreground objects for colour video surveillance. This is presented in Chapter 3.

- Target classification in LWIR video surveillance using contextual knowledge transfer across domains. Output generated from this contextual scheme is presented in Figure 1.5 showing accurate pixel level classifications for input thermal imagery. This is presented in Chapter 4.

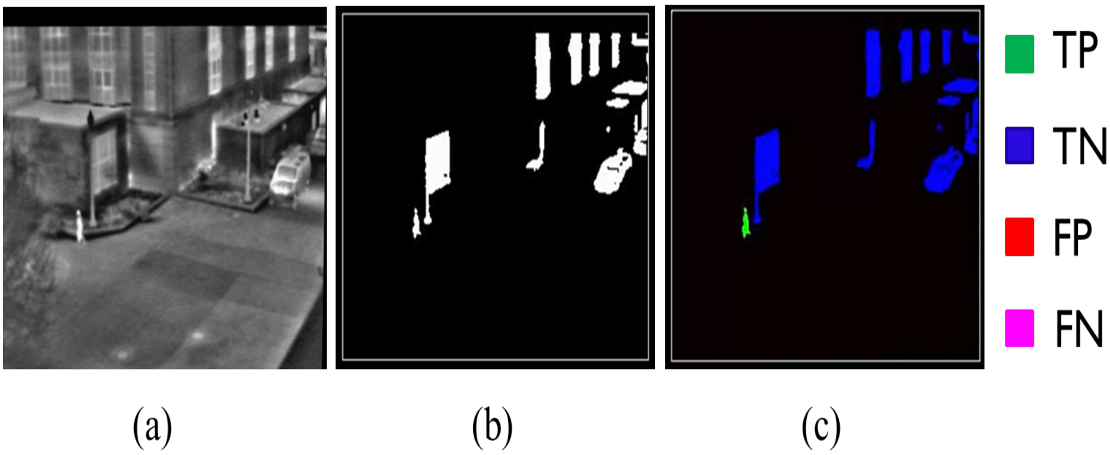


Figure 1.5: Image (a) is a typical surveillance scene captured in LWIR. Image (b) is the thermal signatures our work will classify to provide object information. Image (c) illustrates the output classification from our contextual classification scheme, with the key on the right hand side. The person is clearly identified as green pixels, with background regions shown in blue.

- The design and construction of a deep convolutional neural network for target classification, using LWIR imagery collected via a Thales Catherine MP. The robust algorithm is applied to a variety of scenarios and a realtime demonstrator is implemented to highlight the application benefits. This is described in Chapter 5.

- Long range LWIR target classification is enhanced using scene specific context, for representative problem data collected by a Thales sensor. An illustration of the novel LWIR classifier is shown in Figure 1.4 and the final results are summarised in Figure 1.6, where a significant overall improvement to system performance is gained via exploiting context information for challenging long range data. This is presented in Chapter 6.

1.6 Thesis Outline

The general layout of the remaining thesis follows a certain structure, with regards to the literature review presented in Chapter 2 and the succeeding data chapters. Given the broad range of topics this thesis covers or extends, Chapter 2 discusses related academic and commercial publications to provide a brief overview of the state of the field, for instance identifying deep neural networks as the state-of-the-art method for object recognition, where only *utilised techniques* are reviewed in any great detail. Furthermore, each data chapter contains its own introduction where

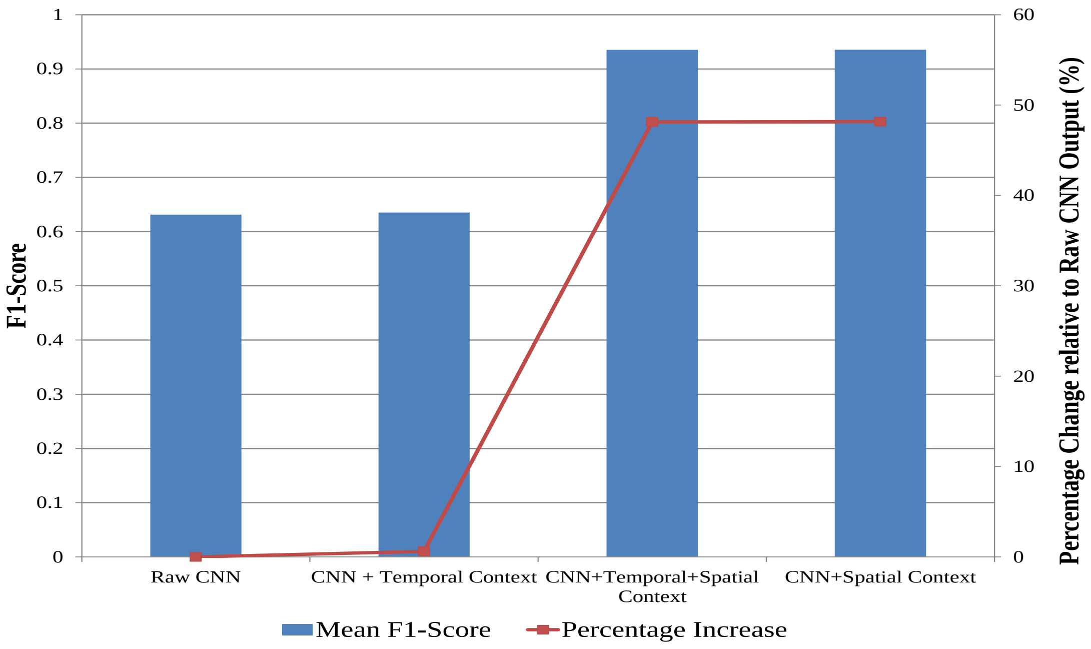


Figure 1.6: The multi-axis plot shows mean $F_1 - Scores$ for the different variants of classification algorithm in our final experiment. The $F_1 - Score$ is a useful summary statistic in machine learning as it provides an a weighted average of a classifiers precision and recall across classes. There is a marked improvement gained from spatial context incorporation. This is highlighted by the red line showing the percentage increase in $F_1 - Score$ relative to the raw CNN output.

similar and relevant works are discussed in more detail to better place the research in context. This structure was chosen for both brevity and clarity.

The data chapters constitute the bulk of the thesis and are outlined as follows. Chapter 3 signals the first data chapter where background recovery of urban scenes is discussed. The principle idea put forth demonstrates that the presence of foreground objects implicitly tells us something about the underlying background in an image, which the method successfully extracts.

Chapter 4 extends the notion of foreground context and background retrieval. This chapter of work incorporates thermal band data and exploits the embedded object information contained within the background recovery scheme, enabling the classification of thermal hotspots *without* a trained classifier. This investigation was completed just as the collective awareness of deep learning and object classification surged dramatically within the field. Chapter 5 discusses the complete process for constructing a target classification algorithm using convolutional neural networks, for LWIR imagery. The source data was collected using the Catherine MP, a high quality TI manufactured by Thales. It is easy to see where the highly accurate output classification scheme could fit into future commercial plans of the company.

In Chapter 6 research elements from the previous three chapters are incorporated into a contextual scene segmentation scheme that can influence the output classification scores of a deep neural network. The method is demonstrated using challenging real world trials data owned by Thales, and is shown to offer considerable performance gains for autonomous long range target classification. Chapter 7 concludes the thesis by highlighting the key points from each chapter and summarising findings. A brief discussion of future work is also included.

Chapter 2
Related Work

Developing algorithms to enhance scene comprehension and autonomous target recognition performance is highly desirable. This goal is currently the focus of both academia and industry, supported by large numbers of publications and heavy investment by commercial enterprise in this area. Such a capability encompasses many important fields within the overall branch of computer vision, which we review in this chapter to identify and address current gaps. Here, we discuss five topics critical to achieving our objective:

1. *A brief overview of the sensor space utilised for all works contained in this thesis;*
2. *Machine learning and its relevance to the encompassed section topics, including how this paradigm is shaping current computer vision research.*
3. *Image segmentation and recent algorithms to parse imagery into contiguous regions;*
4. *The principles of identifying objects of interest within an image and current methods to detect such targets;*
5. *Sources of contextual information and how it can be incorporated with computer vision systems;*

The review is succeeded by a summary conclusion, where we draw together the central themes and relate them to our stated problem of enhanced scene understanding and target recognition.

2.1 Introduction

Humans use the eye as a sensor to collect visual information and effectively understand the surrounding world environment. The ultimate goal of computer vision research is to develop methods that allow computers or machines the capability of sensing surroundings and understanding the data akin to the human visual system. These methods should be capable of recognising patterns within the visual data as well as learn from prior experience to increase robustness and optimise future performance. To give computers this ability is an arduous task for many reasons. Most sensors only record visual data as a 2-dimensional (2D) interpretation. Although alternative methods accounting for an extra dimension are now increasingly commonplace, these practices are relatively in their infancy compared to established sensor data collection techniques. As such this reduction in spatial dimension leads to a corresponding decrease in usable data. Additional elements in visual scenarios such as camera shake, object occlusion and moving cameras add extra complexities to the computer vision problem [25].

The field of computer vision encompasses many smaller points of focus, such as image segmentation, shape description, 3-Dimensional (3D) vision as well as object detection and tracking. There are many more, of course, but this research field is always growing and has recently been enjoying a vast increase in attention. This is partly due to the high demand for computer vision solutions and applications to real world problems. For instance, autonomous vehicle driving and aiding object detection for robotic manufacturing lines [26] [27]. The other major reason it is becoming intensively researched is the recent widespread proliferation of high performance computing ability that allows such systems to exist and the explosion of advanced machine learning techniques [28]. The area of initial concern in our case is the autonomous target recognition scenario using only two sensor modalities, namely a visible and IR waveband sensor. Before reviewing current literature and research trends we begin by examining sensor information and solutions first, as they are an important component to any autonomous vision system.

2.2 Sensor Information

A sensor is a device that can detect an external physical stimuli and reacts to produce a measurable output of some description. The output is very likely to be, especially in the digital age, an electrical signal. There are of course many physical processes that can be detected and measured, for instance numerous biological and chemical reactions that take place in our own body. However, the most important measurable physical phenomena for computer vision applications is the EM spectrum. The bands that are of specific interest here are the visible spectrum, which humans can perceive, and infra-red which lies mostly outside the capabilities of human vision. The range of wavelengths for these bands are approximately $390nm$ - $750nm$ and $750nm$ - $1mm$ respectively [29] [30]. As described in our motivational introduction, Chapter 1.2, we are primarily interested in these bands due to their potential for 24-hour surveillance, improving overall situation awareness. The advantages and limitations for each modality are discussed below.

2.2.1 Visible Band

The visible band is often referred to as RGB , from the Red-Green-Blue additive colour model, or video. The reasoning behind the name *visible* is simply due to humans ability to perceive only this range of EM radiation. The visible sensor mode is common in current day applications, where most smart-phones have an integrated, relatively high-resolution video camera system of some description. Security surveillance also receives significant interest [31], where RGB cameras are the commonly used choice of observation equipment. The standard colour model used for the visible band is the 1931 CIE XYZ Color Space upon which several other colour space models are constructed, such as the *Lab* colour space. This standard represents one of the first attempts at quantitatively describing how humans perceive colour, developed by the International Commission on Illumination (CIE) in 1931 [32, 33].

Another way of describing a visible band sensor, which extends to all sensors, can be obtained if we acknowledge the definition of remote sensing. The technique of obtaining information via a sensor without *in-situ* measurement is described as remote sensing. In other words it is data collection at a distance. Keeping this in mind there are two terms that describe the various types of remote sensors, such as RGB cameras or x-ray medical imaging devices. All types of remote sensors can be divided into either *Active* or *Passive* sensors. The former requires energy to scan and determine properties of a target object, an example of such a system would be a Light Detection and Ranging arrangement (LiDAR) in which a laser beam is required for scanning. A passive system on the other hand can only detect and record energy that is either transmitted or reflected from a target object. Using this terminology a visible band sensor, such as a Closed Circuit Television (CCTV) camera, must be a passive sensor [30]. Additionally, many different types of visible band sensor exist where most are easily accessible. They will vary in specifications as well as price. Thus when choosing a sensor for a particular task, these details will need to be considered to ensure suitability.

One of the obvious advantages gained from employing video cameras as an RGB sensor is the immediate comprehension offered. Given that humans view the world in exactly the same spectrum there is little hassle understanding recorded images, if the data is of sufficient quality. However, there is a major limitation for visible band sensors and that is the variation in objects appearance with changing illumination. Although this fact has been made use of to characterise some objects by their shadow [34], it is still a significant drawback of the visible modality given it becomes almost useless at night-time. To counteract this weakness extra sensors can be implemented to collect EM information in another wavelength range. For some problems a multitude of different sensors are integrated into a system but here, in this case, the discussion is limited to only incorporating an additional IR sensor.

2.2.2 Infrared Band

Where visible sensors fail, infrared can be utilised to overcome their deficiencies. IR sensors, specifically LWIR, do not suffer from illumination issues and are widely used for night-time sensing, hence the potential for 24-hour surveillance. The usefulness of IR based sensing towards low light surveillance is demonstrated in Figure 1.1, but a TI can also offer a distinct advantage for seeing through obscurants such as fog or smoke, as shown in Figure 2.1.

The IR spectrum is large and can be subdivided into a smaller range of bands. While objects above absolute zero in temperature usually emit across the spectrum, sensors are normally engineered to only collect a small portion or range of the wavelength. Several variations and standard subdivisions exist where a governing body sets precedence for their range values. Again the CIE has a reference benchmark for infrared comprised of three divisions. These are simply: IR-A ($0.7\mu m$ - $1.4\mu m$), IR-B ($1.4\mu m$ - $3\mu m$) and IR-C ($3.0\mu m$ - $1000\mu m$). Another scheme is in place by the International Organisation for Standardisation, in addition to an astronomers standard along with various others, where their values differ slightly from the CIE benchmark. However, one of the most useful subdivision schemes in terms of engineering applications splits the spectrum into five components, rather than the common three-way split. This is presented in Table 2.1.

The IR spectrum captures a significant range of wavelengths and this leads to the subdivision schemes discussed previously. Each division shown in Table 2.1 will

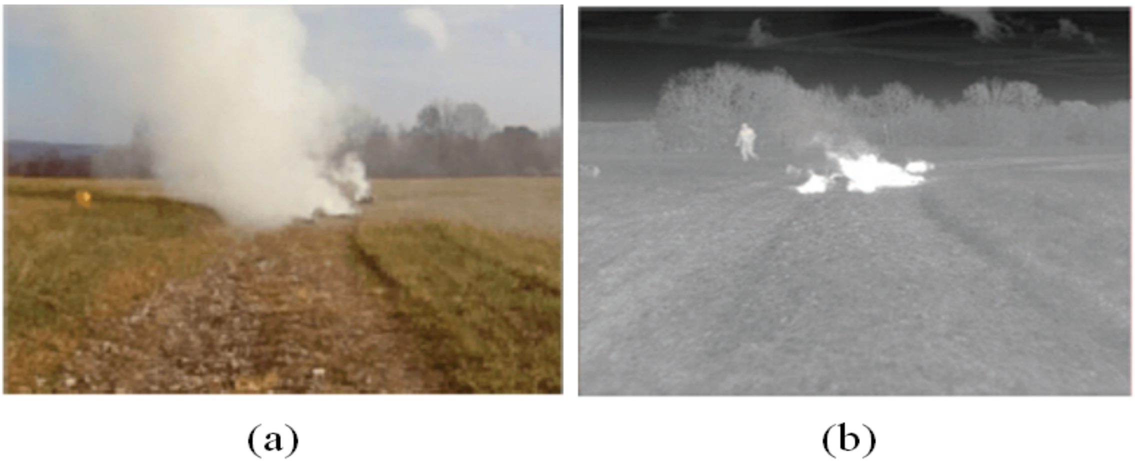


Figure 2.1: Image (a) captures a smoke filled scene using a colour band sensor, where we cannot see anything behind the clutter. Image (b) is a LWIR image of the same scene, where a pedestrian is revealed to be standing behind the column of smoke [1].

Infrared Subdivision Scheme	
Division Name & Abbreviation	≈ Wavelength (μm)
Near Infrared (NIR)	0.75 – 1.4
Short-wave Infrared (SWIR)	1.4 – 3
Mid-wave Infrared (MWIR)	3 – 8
Long-wave Infrared (LWIR)	8 – 15
Far Infrared (FIR)	15 – 1000

Table 2.1: This table illustrates another common infrared subdivision scheme. Thermal IR is usually accredited to either LWIR or MWIR as it is emission dominated. Reflected infrared is referred to as either NIR or SWIR [17].

be suited to particular applications, for instance the MWIR division is used for heat-seeking missile technology [35], which highlights why it is necessary and useful to break up the IR spectrum.

Just like a visible band sensor, an IR sensor can also be described as passive and there are many existing sensor types to choose from. A relevant example would be the Catherine Megapixel (MP) camera produced by Thales. A LWIR TI, it operates over the $8 - 12\mu m$ spectral band. Referring to Table 2.1 it falls into the LWIR category. Additionally Thales also produce a mid-wave version in the same series and this TI is sensitive over the range 3-5μm, falling into the MWIR category in Table 2.1 [36]. Such sensors are not without their own limitations. The obvious issue is trying to detect objects the same temperature as the background, although this is less of a problem for cooled, high quality thermal imagers such as the Catherine MP. Thermal infrared also suffers from a lower signal-to-noise ratio (SNR) as well as limited texture information to help differentiate objects [37]. It is quite apparent now that by using a visible and IR sensor to complement each other, it should lead to a more robust system. However, choosing a suitable sensor is only a small part of designing a solution to a problem. Sensor observations require intelligent analysis and signal processing schemes to transform the signal into a more meaningful representation. One such technique towards obtaining an enhanced understanding of the image signal lies within the field of Artificial Intelligence (AI) .

2.2.3 Data Overview

The most important aspect of building successful computer vision algorithms is access to quality data representative of the problem being solved. To that end, any data used to develop and implement algorithms is explicitly stated in the data chapters. For clarity, however, an overview of available data sources worked on is provided here. For standard colour video surveillance footage CAVIAR [1] and Oxford datasets [12] feature in Chapter 3. These are ground truthed, static camera urban surveillance streams containing pedestrians. A popular colour-thermal registered dataset, OTCBVS [38], is utilised in Chapter 4, which is also static surveillance type footage and ground truthed for tracking pedestrians in an urban environment. This covers all public datasets used to obtain the research results throughout the thesis as all other data is collected and owned by Thales. Lastly, a very recent thermal infrared dataset has been made publicly available by Berg et al. for visual object tracking algorithm evaluation [15].

2.3 Machine Learning

The research domain of AI is vast due to widespread efforts aimed at giving machines the ability to think. The current trend to distinguish between different types of AI is to class methods as either **Strong** or **Weak.** Under this terminology *strong* AI methods relate to machines with human-levels of perception and deduction for a variety of tasks. Conversely, *weak* AI refers to machines with a very narrow or targeted application. The distinction was first coined by the American philosopher John Searle [39]. Presently we are experiencing a strong surge in weak AI methods that are excelling at specific tasks, from mastering the ancient game of Go to large-scale object recognition [40, 41].

[1] http://groups.inf.ed.ac.uk/vision/CAVIAR

Although it is highly desirable to build a true *strong* AI, it remains very far off despite the significant advances made in recent times. One of the challenges related to this is that suitably intelligent algorithms need to extract useful information from raw data to recognise patterns. This process is known as *Machine Learning* and is extremely valuable in modern algorithm design, especially for computer vision applications. The capability to determine meaningful insights from data usually falls under two distinct branches, *Supervised* and *Unsupervised* learning. The latter of these is concerned with drawing inferences from clusters of data that have no associated label or ground truth information. We are not concerned with any such methods in this thesis and shall focus on purely supervised machine learning schemes instead.

An overview of this machine learning subdivision and subsequent classifier construction is presented, along with the classic neural network algorithm. Both of these tools feature prominently in this thesis as well as computer vision in general, playing central rolls in object and scene classification, hence why they are given such attention.

2.3.1 Supervised Learning

The goal of any supervised learning approach is to build a model that maps a number of features to an output prediction. The form of this output can be continuous for a regression process or discrete for classification, distributed over a number of target classes [42]. Supervised methods only require the target label be known during the training stage and are applicable in many circumstances. For discrete outputs, if there are only two target classes it is known as *binary classification*. If there more than two possible target classes the problem can then be described as *multi-class*. In our case we are primarily interested in constructing models with multi-class output so will only consider this option, where object and region types are the most relevant kinds of classification required. We can now delve into an overview of crafting a classification model.

Generally, classification strategies rely on a two stage process. Assuming we have sufficient candidate targets available, the first task is concerned with feature crafting. The aim is to extract relevant attributes from the data that best captures discriminatory aspects of the signal, whilst reducing redundant information. Ideally, these features should be generalisable to the task at hand and offer suitable performance when only a subset of descriptors are available [43]. Extracted features are usually arranged into a vector or $2-$dimensional feature image which ultimately is

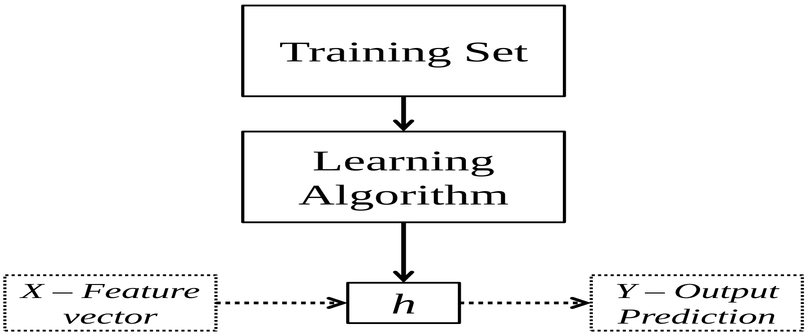


Figure 2.2: Overview of supervised learning process. A functional mapping can be determined via an optimisation process over a training set, which can then transform future inputs to some desired output prediction.

a representation of candidate targets [44].

After successfully assembling a feature construct comes the second stage in the process, requiring a classification scheme. If we let the described feature vector be known as X, composed of n feature instances, then $X = \{x_1, ..., x_n,\}$. Using this terminology allows us to express a candidate target, for example a transformed image blob, as a feature vector X, where the goal in object recognition scenarios would be to determine what object class, C, a potential target belongs to. Furthermore, in the context of machine learning based classifiers an additional choice must be made with regards to the learning paradigm. For such tasks the problem tends to be well bounded and constrained so it is usual to see a supervised learning approach, where labelled training data is required to infer a predictive function.

Thus, a classifier during training would require input pairs of data, formed of the feature vector X and a corresponding label Y. After training is complete a classifier will generate a prediction of Y for any new input feature vector X. The alternative to supervised learning is unsupervised learning, which does not offer predictions of Y for input X but instead clusters the features accordingly. The task of assigning an object class C to candidates based on a feature vector X summarises a target classification procedure [45]. Note that the process remains the same for both regression and classification type problems. As such the entire process can also be shown pictorially in Figure 2.2 where most problems generalise to the approach shown. An example could be to to create a suitable house price predictor, which is the target value Y, based on input features X. Some kind of mapping $h(X)$ from X to Y

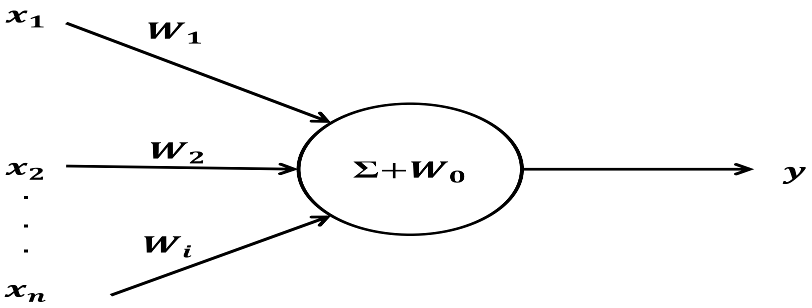


Figure 2.3: Diagrammatic representation of the simple learning unit, the *perceptron*. Input features x form a product with their corresponding weight value w. These are aggregated and a bias weight term is added to give the perceptron response y.

is obtained via the supervised learning process, based on collected and labelled data.

Thus, if we have a training set and labelled targets we need to explore the learning algorithm chosen. Given how vast the field of AI and machine learning is, along with the very wide scope of this project, we shall only review the classic general learning algorithm; the *neural network.*

2.3.2 Neural Networks

A typical Neural Network (NN), also known as an *Artificial Neural Network* (ANN), is composed of many individual processing structures. These unit elements are called neurons, inspired by the human brain, which connect to each other forming a network topology. An individual neuron represents a very simple trainable function, also known as *the perceptron*, capable of performing linear regression and discriminant analysis. An illustration of the unit perceptron is provided in Figure 2.3, showing many input features being fed into the structure resulting in an output response value.

The relationship between input values $x = (x_1, ...x_n)$ and output target y is given by Equation 2.1. Referring to this formula and Figure 2.3, we can observe the main structure being fed with input features, x_i, as neuron or perceptron. The input values are multiplied by corresponding weights, w_i, summed and then added to an additional bias term W_0 to give response value y. The weights can be tuned via a vast array of optimisation strategies, where the goal is to minimise an error function that suits the problem at hand. It is in this way that a single unit can effectively learn a decision boundary or regression line for relatively simple tasks.

$$y = w_0 + \sum_{i=1}^{n} w_i x_i \tag{2.1}$$

Building on this single processing unit, an ANN structure can be assembled by connecting each units into a system. In these interlinked systems unit responses serve as input to more computational processing units further along a network. Input feature values effectively propagate through the NN structure, activating individual neurons in different ways and non-linearities are introduced, which are essential for creating decision boundaries [46, 47]. A *layer* of units represent a computational stage and while the signal activates each neuron individually, responses are aggregated at each layer. Shallower networks have fewer layers while deeper arrangements consist of many and theoretically are more suitable for abstracting better inferences from large, complex datasets.

The architecture of NNs, in terms of neuron capacity, layers and direction of information, can range from simple arrangements to complex structures. Hyper-parameter optimisation schemes are essential to obtain desired performance from NNs regardless of network topology. The classic mechanism implemented to achieve this is gradient descent with sufficient backpropagation of errors, which we will discuss when appropriate in a later section. Furthermore, we shall also cover different network architectures and modern techniques in data chapters corresponding to our ATR solution in the thermal domain. For now, this represents a fair overview of the ANN model. It remains remarkably successful due to its generalisable learning powers. It should be highlighted that while it is theoretically possible for NNs to approximate any function to find a suitable problem solution, the feat remains hard to achieve in practice. There are many factors to consider such as network architecture, initialisation strategy, optimisation scheme and data pre-processing steps to name but a few. All of which play a key role in obtaining an effective prediction model. Overall, neural network models and related variants have been successfully applied to a wide range of challenging tasks, where it appears this status quo will be maintained for the foreseeable future [48].

2.3.3 Conclusion

The aim of this section is not to cover the truly vast subject of Machine Learning in great detail. Instead, its intent is to review fundamental concepts underpinning supervised learning and the key tunable model available in this subject area is discussed. A single unit perceptron is introduced and built up to an ANN. Recent advancements in the field have heavily drawn on neural net models and they also feature prominently in the latter stages of this thesis. Hence they are deserving of our focus and specific architectures for image processing are evaluated in following chapters. The ideas discussed in this section are relevant to image segmentation and object classification. Thus, related machine learning works in these areas are reviewed in their respective sections. We can now examine image segmentation and related methods.

2.4 Image Segmentation

Image Segmentation is a crucial and elemental process in low-level computer vision whereby images can be partitioned or parsed into a set of regions that do not overlap. These regions should be semantically meaningful with respect to a given task and when unified form an entire image [49]. Humans can perform this operation and find it very easy, in most cases, to separate images into contiguous blocks from complex scenes. However, this presents a significant challenge when attempted by a machine [50]. There is a large body of work regarding the problem and as a result, a considerable variety of approaches exist for segmenting images. We shall mostly focus on the key concepts underpinning the topic of image segmentation, with focus given to prominent algorithms and selected recent methods. Machine learning has transformed the field of late and a small discussion is given to this sub-topic towards segmentation.

A wide range of segmentation techniques are available but a robust approach suited to every problem does not exist. Much akin to other sub-topics within the field of computer vision, image segmentation is very application dependent. For instance efficient algorithms for tumour detection in medical imagery [51–53] may not be at all useful for automatic road extraction using satellite imagery [54, 55]. Bearing this in mind we should most likely ignore *objective* segmentation algorithms that are very specific to certain problem domains. Instead we should concentrate on parsing algorithms that tackle *subjective* scene segmentation and are very general in nature, similar to how a human would perform if presented with the task.

Generalised parsing techniques will typically take few inputs to control the granularity and uniformity of output image segments, where contiguous regions are formed from primarily colour and spatial information of the pixels. For example, a user could elect to extract hundreds of very small regions within an image or aim to retrieve a small number of very large segments by tuning the input parameters accordingly. Algorithms adherent to this simple scheme can be generalised to wide variety of problems which is ideal in our case. Recall our aim is trying to improve the comprehension of target scenes, which in turn can be used to improve target recognition and vice versa. Thus the ability to gear a method to whatever scene type is of interest, from *rural* to *urban* environments, is very useful. Note that image segmentation only provides the presence of similar regions in an image, additional steps have to be invoked to provide the scene understanding we wish to exploit. This is known as *semantic segmentation* where regions are provided a class label or meaningful association. The additional steps to achieve semantic meaning for extracted regions are discussed later in this thesis.

Given this overview of segmentation it should now be clear that the field is a wide open problem and application dependent. Ideally an algorithm for our task will be tunable for region extraction and generalisable as a result, despite the contents of target scenery. We will only discuss the mechanism for generating such regions without attached labels for now. Key segmentation techniques fit for this outlined purpose are examined in the sections following.

2.4.1 Graph-Based Segmentation

Efficient Graph-Based (EGB) segmentation is a technique by Felzenszwalb and Huttenlocher published over a decade ago and is a seminal piece of work for image segmentation [2]. Despite the age of this algorithm it remains far from irrelevancy due to its simplicity and unspecific nature, still appearing in recent publications [4, 56]. This is especially impressive in such a fast moving research field. One of the stated aims of the work itself is to provide an effective, general approach to the image segmentation problem domain. The technique also benefits from being rather simple to implement and efficient, resulting in an algorithmic runtime of $O(nlogn)$, for n image pixels. It is a graph-based method and region-segments an image.

The set points of an arbitrary feature space can be represented using a graph theoretical approach. Set points can be represented as a weighted and undirected graph $G = (V, E)$. Each node $v_i \in V$ corresponds to a point in the feature space, an image pixel for instance, where the edges $(v_i, v_j) \in E$ connect neighbouring pixel pairs. Pixels are related using the 8-connectivity scheme and pixel connections are used to construct the edge set E. Furthermore, an associated weight $w(i, j)$ is given to each edge and is a function of the similarity between nodes i and j. Weights are then given as $w(v_i, v_j)$ and can be calculated as the absolute intensity difference between edge pixels. This is shown in Equation 2.2 below.

$$w(v_i, v_j) = |I(p_i) - I(p_j)| \tag{2.2}$$

Here, $I(p_i)$ is intensity for the pixel p_i.

Now for a graph based approach, a segmentation S partitions the vertices V (recall V is pixels) into regions/components. Each component $C \in S$ should then correspond to a connected region in graph $G^{'} = (V, E^{'})$, such that $E^{'} \subseteq E$. To elaborate further, an image segmentation is a subset of edges present in E. The desirable outcome for a segmented image is that similar portions within the image are grouped together and separated from dissimilar portions. Using the terminology outlined above, this relates to edges between vertices within the same portion having low weights. The opposite is also expected, where edges between vertices within different components exhibit high value weights. Overall, this results in an image being partitioned into spatial regions of similar pixel intensities. Once we examine this method in detail, it is clear some kind of mechanism must exist for determining if a boundary should exist between two portions in an image segmentation.

Consequently, this is achieved via calculating the dissimilarity between elements along the components boundary, relative to the dissimilarity of collective pixels within each component. This can be summarised as a comparison of *inter* and *intra* component differences. An internal difference of $C \in V$ as the largest weight in the Minimum Spanning Tree (MST) of component, $MST(C, E)$ is then defined, determined by Equation 2.3.

$$Int(C) = \max_{e \in MST(C,E)} w(e) \tag{2.3}$$

The underlying reasoning behind this measure is that a component C only remains connected when edges of weight *at least* $Int(C)$ are considered. It follows that the minimum internal difference, M_{Int}, can be described as shown in Equation 2.4.

$$M_{Int}(C_1, C_2) = min(Int(C_1) + \tau(C_1), Int(C_2) + \tau(C_2)) \tag{2.4}$$

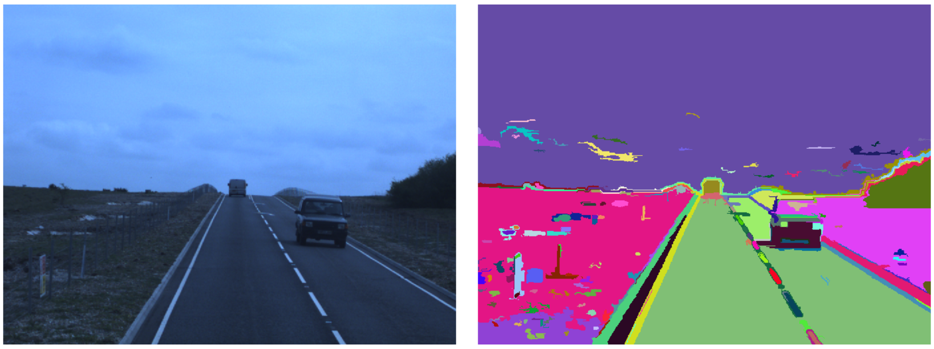
Here the threshold function τ acts as a sensitivity control for inter-component difference *relative* to intra-component differences, such that the inter region difference must be greater to present evidence of a boundary between them. This thresholding function is based on component dimensions and can be simply expressed as $\tau(C) = k/|C|$. Here $|C|$ is the size of component C and k is an some arbitrary constant parameter. In other words, smaller components essentially need more evidence for a boundary to exist. Moreover, k sets scale size where a greater k tends to result in larger image segments and lesser values of k produce smaller image components.

The algorithmic steps for EGB segmentation technique can now be summarised. Given an input graph $G = (V, E)$, with n vertices and m edges, the desired output is a segmentation of V into components $S = (C_1, \ldots, C_r)$. The steps are then as follows in Algorithm 2.1:

Algorithm 2.1 Efficient Graph based Segmentation

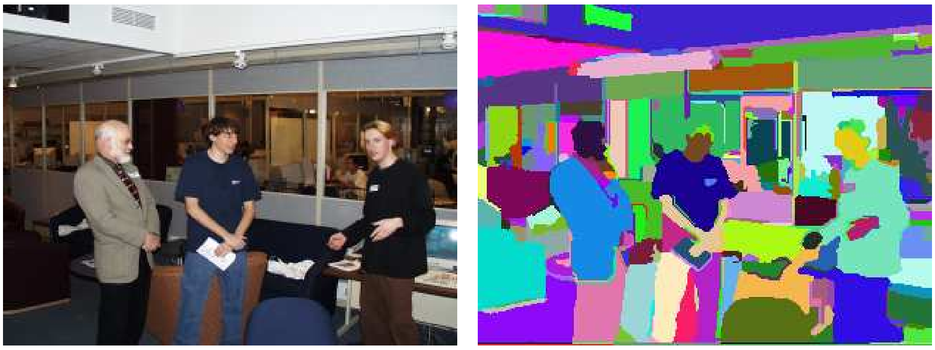
```
0 Sort $E$ into $\pi = (o_1, \ldots, o_m)$, by non-decreasing weight.
1 Start with a segmentation $S^0$, where each vertex $v_i$ is in its own component.
2 Repeat step 3 for $q = 1, \ldots, m$.
3 Construct $S^q$ given $S^{q-1}$ as follows. Let $v_i$ and $v_j$ denote the vertices connected by the q-th edge in the ordering, i.e., $o_q = (v_i, v_j)$. If $v_i$ and $v_j$ are in disjoint components of $S^{q-1}$ and $w(o_q)$ is small compared to the internal difference of both those components, then merge. Otherwise do nothing. More formally, let $C_i^{q-1}$ be the component of $S^{q-1}$ containing $v_i$ and $C_j^{q-1}$ the component containing $v_j$. If $C_i^{q-1} \neq C_j^{q-1}$ and $w(o_q) \leq MInt(C_i^{q-1}, C_j^{q-1})$ then $S^q$ is obtained from $S^{q-1}$ by merging $C_i^{q-1}$ and $C_j^{q-1}$. Otherwise $S^q = S^{q-1}$.
4 Return $S = S^m$
```

Although the EGB algorithm is outlined there are still some further points to consider. Before edge weights are computed for an input image, a Gaussian filter is applied to smooth out any present image artefacts. This is one of three user controlled parameters where the degree of Gaussian smoothing is chosen by σ. Note the remaining parameters are k, the aforementioned threshold function constant, and *min*, which enforces the minimum component size in post processing. Also, for colour images the algorithm separates input structures into three monochrome planes (RGB) and runs once for each. Final regions are determined by intersections of these and finding common components. Example images processed via this technique are presented in Figure 2.4 and Figure 2.5.



(a) Input Image (b) Graph Based Segmentation

Figure 2.4: Using a frame from a Thales data gathering project, Hydravision, the effects of EGB segmentation are easily observable. The input image shows a typical rural road scene and the overall segmentation is sensible - although some bleeding can be seen on the right hand side as the grass banking merges into the windscreen of oncoming traffic.



(a) Input Image (b) Graph Based Segmentation

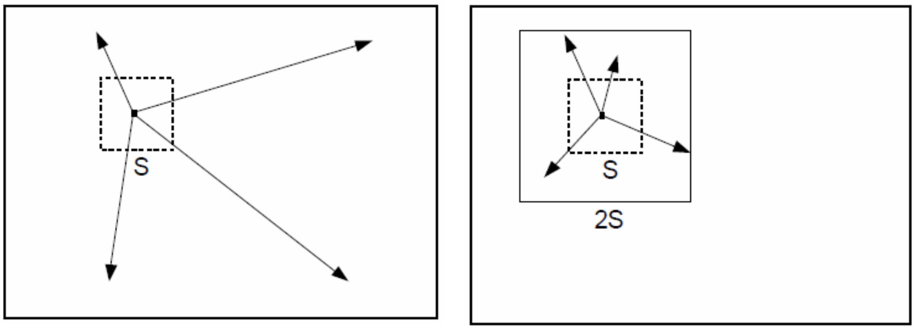
Figure 2.5: Another example of this segmentation scheme working on typical office scene. It appears reasonable given the input image and some contiguous regions have formed. Image source [2]

In general, this segmentation algorithm offers reasonable performance but suffers from a variety of issues. For instance the segmentations are prone to bleeding, as highlighted in Figure 2.4. This is where different image segments incorrectly merge into neighbouring regions. Furthermore, the algorithm can be rendered useless by choosing bad parameters, given the delicate tuning mechanism it can be easy to observe this behaviour. Lastly, the segments can be highly irregular in shape but this may not be a hindrance depending on the desired application. Other than these drawbacks, which are mostly related to parameter selection, the method is efficient and generalisable. It still holds up well to recent segmentation algorithms, compared via benchmark evaluation metrics, while still garnering citations from methods incorporating the basic function of EGB segmentation. One recent trend in this general image parsing domain has been that of the *superpixel*. This is discussed in the following subsection.

2.4.2 Superpixels

Superpixels are simply smaller segments of an overall image region. They are constructed by grouping picture elements, i.e. pixels. A variety of techniques now exist to accomplish this feat [57–59]. A general rule of thumb is that grouped pixels will all share similar properties whether it be spatial or intensity values. The motivation for using superpixels instead of individual pixels was first targeted by Ren and Malik [60]. They illustrate generating optimal solutions at the pixel level is demanding because there are so many pixels at moderate image resolutions and above. By grouping elements into larger, atomic regions capturing key image structures, it can effectively reduce the complexity of any further image processing operations. In essence, superpixel methods are simply restricted region segmentations, although the regions produced from the EGB algorithm discussed previously can also be regarded as superpixels [61].

At the outset of this EngD project the recognised state-of-the-art superpixel method was considered to be the Simple Linear Iterative Clustering (SLIC) algorithm [3]. It still remains a hugely popular choice and we elect to examine it in detail given it features prominently in our early research. The SLIC algorithm is a gradient-ascent based method. An initial rough cluster of pixels is the starting point, which is iteratively refined until a convergence condition is fulfilled. Reaching this point results in the formation of superpixels. Note this differs from the other category of superpixels, namely graph-based techniques such as that of [2]. For graph-based superpixel methods, each pixel is a node in a graph. Edge weights



(a) Classic k-means search pattern (b) SLIC reduced search area

Figure 2.6: Image (a) demonstrates how regular k-means algorithms search the entire image field. Image (b) highlights the constrained SLIC search area leading to a more efficient algorithm [3].

are then assigned between nodes according to similarity measures and superpixel formation is obtained via cost function minimisation over the graph.

The SLIC algorithm is an adapted k-means clustering approach for constructing superpixels. It employs two key adaptations leading to a reduction in algorithm complexity and improved superpixel properties. The modifications are limiting algorithmic search space and using a weighted distance measure to incorporate colour-spatial properties, respectively. An illustration of the constrained search space process is presented in Figure 2.6, originally described in [3].

For input colour imagery SLIC operates in the *CIELAB* colour-space [62]. Initial user defined parameters indicate how many superpixels are desired, given as k. This allows k cluster centres to be sampled over the image grid, spaced S pixels apart. This spacing is designed to produce roughly equal sized superpixels and can be calculated by $S = \sqrt{N/k}$, where N is total number of pixels. Cluster centres are readjusted to the lowest gradient position in a 3×3 neighbourhood, which has the benefit of avoiding placing a centre on an edge or noisy pixel. Pixels are assigned to the nearest cluster centre whose search region overlaps its own location. This also leads to an algorithmic speed-up due to reduced distance calculations. The expected spatial extent of each superpixel is given as $S \times S$, shown in Figure 2.6, where the whole search region for similar pixels is given as $2S \times 2S$. After pixel-cluster association, the centres are adjusted to the mean colour and location vector $[labxy]^T$. A residual error is calculated iteratively until convergence and lastly, rogue dis-joint pixels are assigned to the nearest superpixel to enforce connectivity.

The SLIC algorithm creates superpixels corresponding to clusters in the *labxy* colour space. As such, the distance measure D has to be carefully designed. Firstly, D computes the distance between a pixel i and the cluster centre C_k. Pixel colour information is in the CIELAB colour-space $[lab]^T$, where the range of possible values is known. However, pixel location given by $[xy]^T$ may take on a varied range of values depending on the image dimensions. Thus, by only defining the distance measure D in the *labxy* space alone will result in unpredictable clustering behaviour for different sizes of superpixel. For instance, larger superpixels will be dominated by spatial distances relative to colour proximity, generating superpixels that do not adhere well to image boundaries, which is often a target evaluation metric for segmentation tasks. It is obvious that the reverse is also true for small scale superpixels.

To overcome this problem a combined distance measure is introduced, the function of which is to normalise spatial and colour proximity by their respective intra-cluster maximum distances [3]. These are given in Equation 2.5, where m is a constant that controls relative superpixel *compactness*.

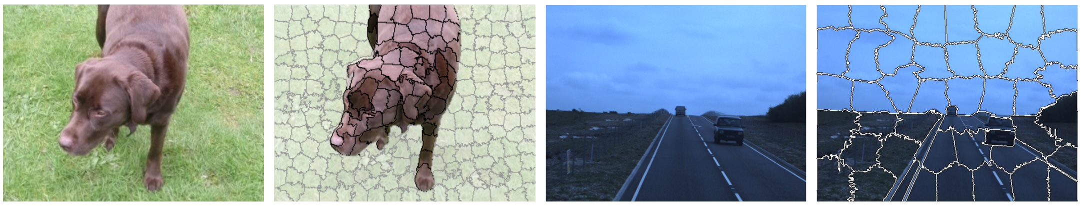
$$\begin{aligned} d_c &= \sqrt{(l_j - l_i)^2 + (a_j - a_i)^2 + (b_j - b_i)^2} \\ d_s &= \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} \end{aligned} \tag{2.5}$$

In the original implementation of the SLIC algorithm, m can be in the range $[1, 40]$. For large values spatial proximity is regarded as more important leading to compact superpixels. Conversely, low values for m lead to superpixels that adhere well to image boundaries, but have less regular shapes. It should also be pointed out that the algorithm can handle grayscale images by setting d_c to only use the first squared term. These are then combined and simplified into the final distance measure provided as Equation 2.6, utilised throughout the algorithm.

$$D = \sqrt{d_c^2 + (\frac{d_s}{S})^2 m^2} \tag{2.6}$$

Example images are presented in Figure 2.7 illustrating how SLIC transforms the image representation. Performance metrics reported in the work also highlight the algorithms position as state-of-the-art in terms of segmentation evaluation at the time of publication. Generally, this superpixel method performs well and is flexible when utilised.

Having covered the SLIC algorithm in detail, as well as the EGB segmentation technique, we should now consider where the machine learning driven SOA lies.



(a) Input image 1 (b) SLIC processed version 1 (c) Input image 2 (d) SLIC processed version 2

Figure 2.7: Image (a) presents a dog from the PASCAL dataset [63] and serves as input to the SLIC algorithm. An output superpixel representation is generated using this algorithm and is shown in image (b). Images (c) + (d) follow this same process, where (c) is a frame from the Thales data acquisition project. Algorithm parameters were slightly different for each input image to illustrate a difference. This algorithm appears to adhere very well to boundaries in an image.

2.4.3 Segmentation & Machine Learning

So far we have only surveyed the principle of parsing an image into regions, without any knowledge of what the image segments could represent. One of the only assumptions we can make about the contiguous regions is they contain similar pixel values and are fairly homogeneous in nature. This tells us nothing about what is contained within the parsed scene which is an important piece of information, especially in situational awareness terms. Scene structure and objects within a scene tend to have a well-defined relationship, as we shall explore in contextual terms, so being able to determine region classes is highly desirable.

Fortunately, machine learning provides many ample tools allowing region classification. Of course in a supervised learning setting the caveat requirement is labelled segmentation data. We should note that machine learning methods are not just limited to providing semantic meaning to segments, they can actually improve output boundary adherence [60]. However, we are only considering the region classification capability offered via supervised learning methods.

There has been a recent surge in techniques utilising neural networks towards achieving robust semantic segmentation, specifically Convolutional Neural Networks (CNN) which are a special network architecture designed around the $2-$dimensional structure of imagery. One of the most well-known of these methods is presented by Zheng et al. [64], using Conditional Random Fields (CRF) as Recurrent Neural Networks (RNN) . In this prior work the authors employ a complex deep network structure that exploits the advantages of both CNNs and CRFs, allowing accurate fine grain pixel labelling of scenes. The model can be trained in an end-to-end

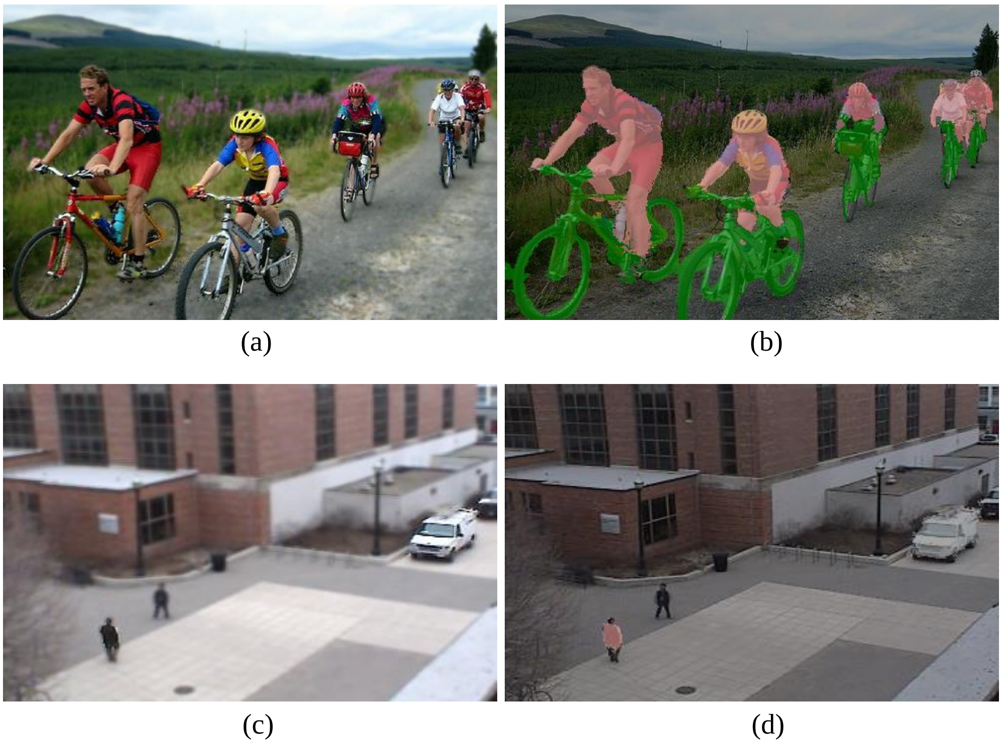


Figure 2.8: The quadrant of images shows the CRF as RNN algorithm performing semantic segmentation on wo sets of images. Image (a) is an input example of humans riding bicycles in a country scene. Image (b) shows the result of the semantic segmentation with bicycles and humans being mostly covered in green and pink overlay respectively. Image (c) is a different input scene, more representative of a surveillance type scenario, showing a sparsely populated courtyard. image (d) shows the resulting semantic segmentation, with one person being highlighted in pink and a van in the top right overlaid with white cover.

fashion once assembled which is especially advantageous. An example semantic segmentation using this technique is illustrated in Figure 2.8.

The work of [64] is just one recent example of an accurate semantic segmentation scheme. There are an increasing number of alternative methods aiming to achieve pixel-wise labelled segmentation output, similar to the output shown in Figure 2.8. While the accuracy and efficiency of competing methods is at least similar or improving, they all share a common utilisation of deep neural networks to achieve their goals as it enables much better inferences to be drawn [65–69]. However, if we observe the scenes shown in Figure 2.8 it is immediately apparent that only objects within the scene have been extracted and labelled. This is almost the wrong way round from our desired behaviour, where we want to segment background regions and identify the semantic class of these instead. This is impossible with the

CRF-RNN network for example, as it is trained only on what we would describe as *foreground* objects such as person, bicycle, car, or horse. We propose that foreground objects exist on the underlying *background* of scenes, such as a pedestrian walking on road then grass. The walker would be the foreground object and the road or grass would be deemed a background region. This idea will be revisited in great detail later.

Despite the ever increasing performance of such semantic segmentation methods focused on foreground objects, they will not be effective for extracting and classifying background regions. Referring to the growing list of semantic segmentation techniques, it is easy to observe the current trend for machine learning applied to image segmentation methods. Currently the core idea in the field is to develop segmentation models that adhere very closely to foreground object boundaries, followed by extraction and classification. This is exactly what we illustrated in Figure 2.8. Most methods are geared towards finding foreground objects to the exclusion of the underlying scene background and they are beginning to show robust performance. It is a perfectly understandable route to undertake as objects appearing in scenes are very likely to be the most important *thing* there. This strand of research is built on the principle that foreground objects exist on background regions, where underlying segments need to be mostly excluded or filtered out before classification.

Now, what if we apply one of these advanced techniques to a scene typical in the security and defence domain? In all likelihood the algorithm will return incorrect or no extracted foreground objects, especially in long range rural environments where targets are small and inconspicuous. This is because objects of interest in defence scenarios will either be sparsely populated, actively trying to avoid detection or at considerable distance from the collecting sensor. All the semantic segmentation methods highlighted above are trained to handle much more sterile imagery usually found on public datasets. As such it is very likely they could not handle the challenging imagery we have to address in this research project. To tackle the outlined problem we approach semantic segmentation by acknowledging the background regions that foreground objects inhabit.

There are few related works that adopt this perspective for semantic segmentation, where we aim to classify the background regions themselves to aid target classification. Two recent methods utilise EGB segmentation and a region classifier to obtain improved scene context [4, 70] and we employ key ideas from these works in Chapter 6. Both of these techniques present general segmentation methods capable of extracting and classifying regions such as sky, water and grass. An example

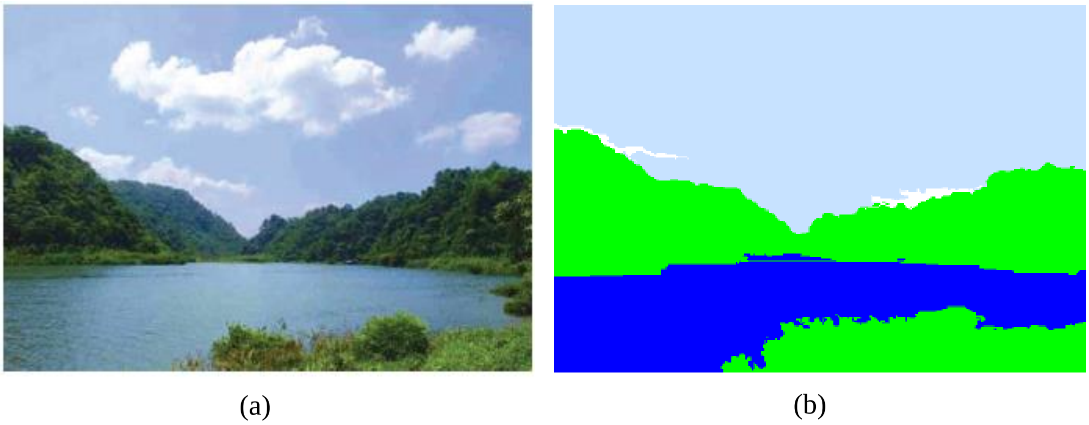


Figure 2.9: Demonstration of region segmentation and classification. Image (a) is the input and (b) is the labelled output, where water is dark blue, grass is green, the sky is light blue and white pixels are unclassified. Image source [4].

image from [4] is presented in Figure 2.9 highlighting a classic region segmentation.

Ultimately, the foundational idea for this approach acknowledges the tightly coupled relationship between objects and backgrounds. Foreground *things* exist and behave in specific, structured ways depending on the surrounding environment. This knowledge can be exploited to tell us what kind of objects would be more likely to appear, or not, given the underlying region information in a target scene. For instance, if you extract a road region it is more likely to observe a car than a helicopter occupying the region. While recent semantic segmentation techniques are good at segmenting and classifying objects, it is perhaps detrimental to decouple regions from foreground objects. This realisation is a powerful idea and one that we show to be beneficial for target recognition in challenging environments. Any region classification still relies on supervised learning techniques as discussed previously, but the additional scene perception it enables is significant.

2.4.4 Discussion

The methods discussed only present a snapshot of available segmentation techniques, selected due to their generalisable approach which we leverage towards our problem. Both methods covered in detail feature prominently in early data chapters, with EBG making a further appearance in the final research chapter. They should highlight the different approaches to segmentation but also bring to focus that there is no obvious answer to image segmentation. When considering a specific application, it does indeed narrow down the field of available techniques but the remaining methods are vast and varied, not only in methodology but also in evaluation metrics.

We also reviewed recent advancements in semantic segmentation to gain enhanced scene understanding. Such methods rely on machine learning in conjunction with existing segmentation algorithms. However, the majority of these works will not be suitable for challenging defence and security based imagery that we essentially need to work with. As such, we find some of the few relevant works using region based classification and frame the problem in a new way, where foreground and background information is tightly coupled and mutually beneficial to identify. We can now examine relevant object detection and classification techniques.

2.5 Object Detection & Classification

The research field concerning object detection and classification is a wide and varied one, encapsulating several fundamental problems within computer vision. Recent advances in this field have been notable, especially on the classification front, due to improved machine learning methods being incorporated. Features representations of target objects are often combined with deep learning methodologies to create robust object detectors that also provide output classification scores. The norm until roughly five years ago was for a human to painstakingly hand-craft these feature representations, but now we can simply give a machine lots of examples to generate these ideal features automatically. We shall explore this concept later but before delving into the mechanics of object detection systems, we must first ask what exactly is an object?

Generally speaking an object is a well defined and independent *thing*. Using computer science terminology we can describe objects as an instance of a class. For example, a *dog* is an instance of the class *animal* etc. In computer vision it can also be said an object should have a well defined boundary and not blend with the incoherent background. Furthermore, it is worth noting the tight bounding of object detection and recognition. As we shall see these two tasks are very similar and increasingly reliant on machine learning methods for feature generation. Given this generic overview of what an object actually is we can move on to discern the motivation for creating detection systems.

The field of object detection remains heavily researched due to the enormity of potential applications, such as autonomous vehicles or improved border protection. Whilst the human vision system is what computer vision research aspires to recreate, our processing capability is limited. Humans are also susceptible to mistakes and

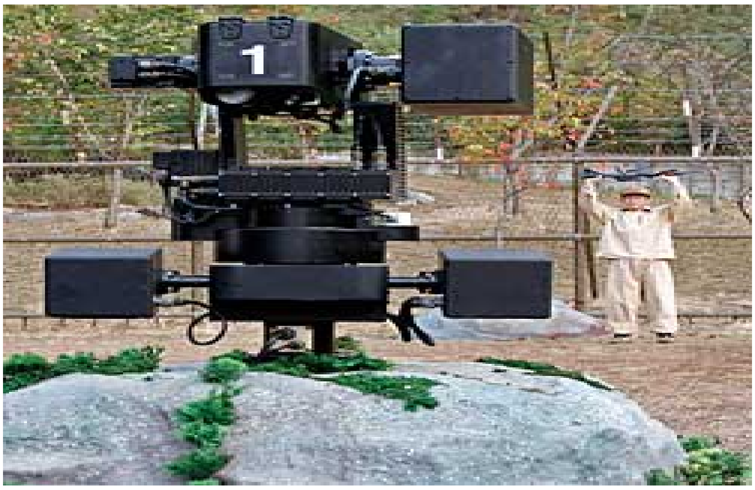


Figure 2.10: A robotic sentry gun developed by Samsung Techwin Co. is shown holding up a potential threat. The SGR-A1 robot uses pattern recognition and object detection based algorithms to identify targets, effectively filtering friend from foe. This is just one such example of object detection being applied to real world technology. Image source [5]

under-performing in pressured environments. This can be a critical hindrance in tasks where fast and accurate information processing is required to make correct decisions. Any number of situational awareness tasks could provide one such scenario, for instance surveillance over the de-militarised zone (DMZ) at the North/South Korean border. Without discussing the details of why this area is of significance, the topic does provide an example of a real-time detection system where situational awareness is of utmost importance. The South Korean military utilise a specially designed stationary robot to over-watch the DMZ. Samsung manufacture the SGR-A1 system and an illustration is presented in Figure 2.10.

The SGR-A1 is a real time object detection/recognition system capable of distinguishing between non-targets and targets. Its sensors operate in the thermal and visible bands. This example highlights the application of object detection systems in a military and robotics domain. Other areas where object detection can be applied is automotive safety, industrial inspection and elderly care. Overall, object detection systems usually focus on pedestrian detection initially and expand to include other objects such as vehicles. This is in part due to the goal of getting machines to successfully interact with humans and our environment, where potential benefits would be enormous [71], but also because it constrains the problem and allows better progress to be achieved.

The rest of this section will explore classifier design applied to the detection problem. We shall then review some of the key papers in object detection over the last decade before finally examining the very recent state of the art in this area. We do this as it is important to highlight the step-change that has occurred in computer vision over the last few years with the mass adoption of machine learning techniques. Progressing from human designed features with simple classification methods to full autonomous feature extraction with much better performance, in the space of a decade, is quite remarkable. The older works and related publications remained very popular even at the outset of this project and feature in the early data chapters following this. Thus, they are included for relevance and not solely for posterity.

2.5.1 Designing A Classifier

There are many algorithms and schemes designed towards object detection [6, 8, 72–77]. In essence they all aim to do the same thing; create a higher level understanding of a scene given the low level pixel information. This understanding could be location or object class for instance. Furthermore, every method is subject to overcoming the same challenges such as dynamic illumination, varying pose, object articulations, occlusions and clutter. Algorithms built to this purpose will usually follow a three part scheme, where the first step is to define and learn ideal features of the target object followed by detection of these objects at test time. The final stage relates to rejection or acceptance of object appearance. This general formula is observed in many prior works in this area and is shown in Figure 2.11. We can now discuss features in more detail.

Features are properties that describe a class very well, in the case of an image object this could be things such as intensity, colour or gradient information. For the application of object detection, features are used to build a classifier capable of indicating whether a given image contains a target object or otherwise. An ideal feature should be local and invariant, meaning it is robust to occlusions and unaltered after a transformation. It should also be robust to noise, blurs and compression as well as highly distinctive, in that individual features can be matched to a large database of objects. Lastly, it would also be ideal if the feature was computationally cheap so real time deployment remains feasible. Given these desirable feature properties it should be immediately obvious that not all will be attainable for the vast majority of cases.

One of the most powerful and conceptually simple detection techniques is generally known as the *sliding window method*. It involves building a dataset of labelled images containing positive and negative instances of the target objects, where neg-

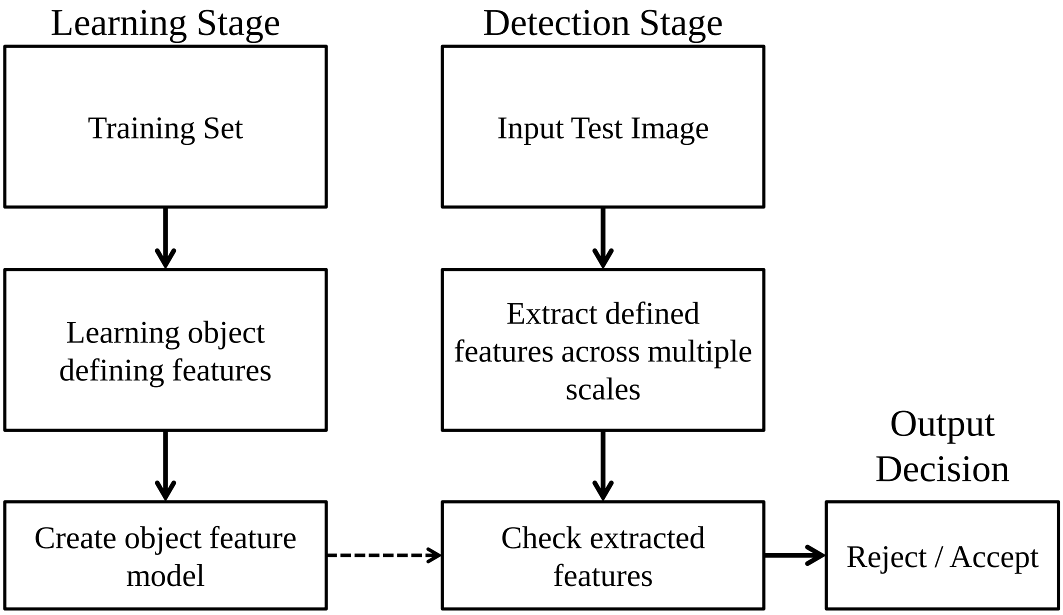


Figure 2.11: A generalised overview of the classical process for creating an object detection algorithm. It is summarised as a three stage process. The training stage involves feature learning to maximise discriminative power. The detection stage effectively deploys the algorithm and tests against input imagery, where the final stage is an output decision on whether an object instance has been detected or not.

atives do not contain the object of interest. A classifier can then be trained in a supervised learning fashion, allowing the determination of correct object appearance. Once this is complete a sub-window of fixed size can be passed across the image, where classifier application returns positive and negative results for each step. This is a search over location and there are two subtleties to this approach.

Firstly, the target objects will be present at different scales within the image, so a search over multiple feature scales also needs to be performed. Secondly, as a sub-window moves closer to a target object it will return more results than necessary, this will skew the algorithms true accuracy and has to be accounted for by Non-Maximum Suppression (NMS) . In other words, NMS performs a local maximum search where only one result is returned for each object instance. The output return is also a higher response than all neighbouring results. There are several variations of NMS that can be applied, such as $1D$ straightforward NMS, $1D$ dynamic block or

$2D$ stripe NMS. Each will have their own nuances and should be carefully researched before applying [78].

The sliding window method is only one such approach for object detection and it is not free from performance issues. For instance, sliding windows do not deal well with occlusions of target objects. The method is also heavily reliant on the assumption that target objects remain fairly rigid and non-deformable. However, even with such drawbacks the concept is very easy to understand and apply. This is a driving force behind its popularity and it seems to compare well with more sophisticated detection methods like *latent-SVM* [77]. Given the impact of sliding window approaches in the object detection field over the past decade, related methods provide the focus of attention in the sections to follow. Having discussed classifier creation and the sliding window approach for detection we can now observe the key ideas populating this subject area.

2.5.2 Classic Object Detection

The feature descriptors known as Histogram of Oriented Gradients (HOG) presented by Dalal & Triggs and the object detection framework outlined by Viola & Jones, probably represent the two highest impact publications in this area over the last decade [6, 74]. At the time of writing, the 2001 paper from Viola-Jones had ∼14570 citations and the 2005 work from Dalal-Triggs had ∼17475 citations, such is the influence of these works. Although the methods are somewhat dated it is worthwhile discussing their main characteristics, especially as a few aspects are still exploited in detection systems. They are briefly discussed in the following sections.

The Viola-Jones framework relies on representing images in a way that allows for sufficiently fast feature computation, an idea coined *Integral Images*. It is an intermediate image representation containing the sums of neighbouring pixels at each image pixel location. A set of rectangular features similar to Haar basis functions is also incorporated into the object detection system. These features can be computed rapidly at many scales thanks to the integral image construction, some example rectangle features are illustrated in Figure 2.12 below. Objects within images are then classified using simple feature values. Utilising features in this way, along with additional machine learning techniques, provided Viola-Jones with (then) state-of-the-art detection rates. Arguably, the most important point from this work is to scale the features and *not* the test images when deployed.

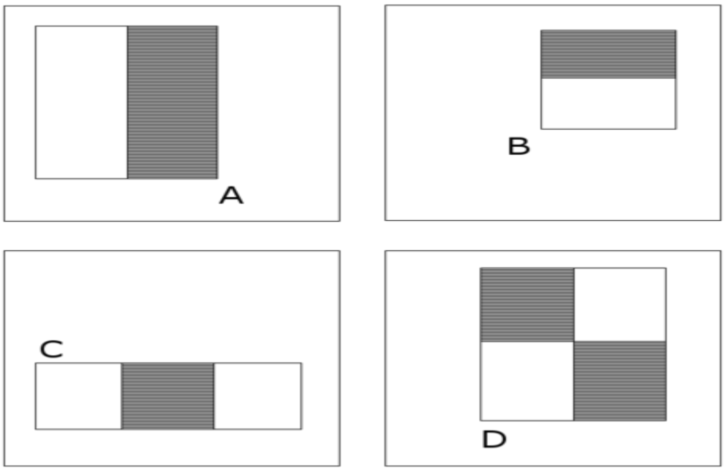


Figure 2.12: These example rectangle features are utilised in the Viola-Jones framework. Their simplicity allows fast computation which is desirable if real-time deployment is a goal. Image source [6].

Following this Dalal and Triggs developed the now seminal work on object detection, creating feature descriptors known as HOG. The rationale was to devise a technique capable of handling the multiple poses of humans for pedestrian detection tasks. This led to the robust HOG feature set, where the frequency of gradient orientations are tallied in distinct portions of an image. Machine learning techniques, such as Support Vector Machines (SVM) for classification purposes, are also applied resulting in am effective method of detecting pedestrians.

Whilst these two works provided massive improvements in terms of robust features, efficiency and applying machine learning methods they still suffer from many issues. Things such as partial occlusions or changing illumination will degrade detection accuracy, but these problems are applicable to the majority of algorithms. The real underlying problem with these detection techniques lies in the feature space; they are simply not good enough. The easiest way to observe this claim is by considering the image presented in Figure 2.13. It illustrates a high confidence detection for a car using HOG features and SVM classifier. It is clearly not a vehicle contained within the sliding window, but a textured water region. This is inherently obvious to a human but not so for a machine. The incorrect decision is explainable upon a closer exploration of the features themselves, where it becomes clear the body of water does indeed look like a vehicle in HOG feature space. This serves as a good reminder that machines can only deal with the information passed to them. The feature *inversion* highlighting this point is illustrated in Figure 2.14.

It is clear that features such as HOG are not good enough in isolation. As robust and simple as they are it still appears as the incorrect class of object when

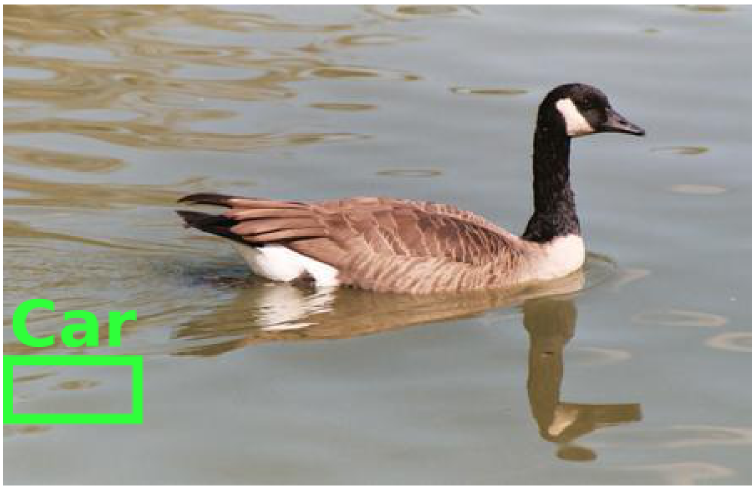


Figure 2.13: An example image of a swan on a body of water. A deformable parts model vehicle detector using a HOG feature space is passed over the image. Green bounding box indicates presence of a car at that location in the image. It is obviously an incorrect detection for a human visual system but understandably incorrect in terms of machine vision [7].

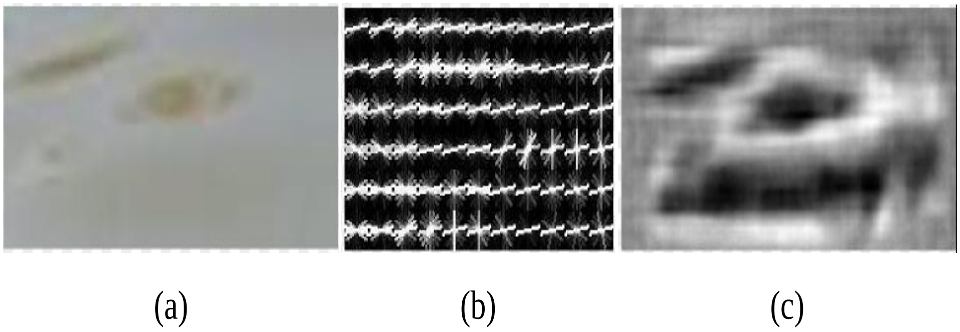


Figure 2.14: Image (a) is water cropped from the sliding window shown in Figure 2.14. Image (b) is the generated HOG features from the water cropped image and image (c) is an inversion of these features. Image (c) effectively illustrates that the water image does share similarities with a vehicle in HOG feature space. Images from [7]

.

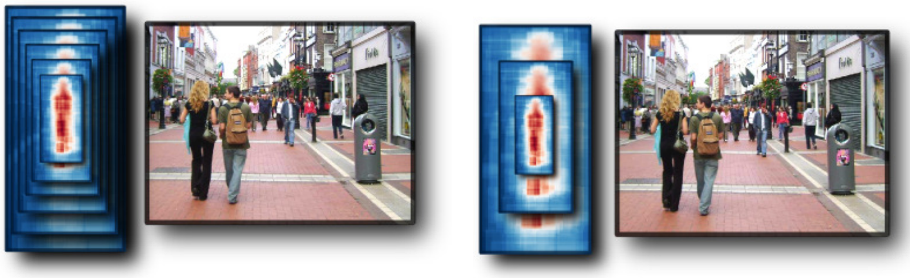
transformed into feature space. Output responses will be high in confidence value but incorrect, generating a false positive The over-arching view of this problem is that machines are only as good as the information it is given, they do not possess a higher level understanding of the world to make decisions as easily as a human would. Recent work on detection algorithms has realised this problem with features and seemed to have added more information to the older methodologies of HOG and Viola-Jones. For instance, the principles of HOG-like features and scaling are maintained in recent work by Benenenson et al., creating a detector named *VeryFast* [8]. This system was recognised as one of the benchmark pedestrian detectors in terms of accuracy and most certainly speed, at least for a sliding window approach.

2.5.3 Multi-Scale Detections

The *VeryFast* detector by Benenson et al provides enhanced detection accuracies over preceding methods, along with phenomenal speed of execution. Again, it is first and foremost a pedestrian detector but the principles can be extended to detect other target objects. The improvements of VeryFast are achieved by two key contributions, namely the better handling of feature scales and exploiting stereo images to create what is known as a *stixel world*. This stixel world essentially exploits multi-camera viewpoint to obtain depth information, which can be used to enhance the entire system. However, it is the treatment of features over scales that is interesting and remains relevant.

The method employed for handling the scales of image features also provides an additional speed-up and helps avoid troublesome tuning issues. The typical approach to object detection for class specific detectors is the sliding window method, whereby classifiers for each position and scale compete against each other. This is not an ideal approach as it requires a high number of models *N*, where a representative value quoted in related literature is ~ 50. This means the model training stage is highly time consuming. The other common approach is to create and train a single model that should, in theory, work for multiple image scales and thus different sized objects. Then the image is rescaled multiple times to fit the single model.

Again, this approach is fraught with performance issues where the dominating factor is resizing input images *N* times when deployed, as well as additionally recomputing the image features *N* times. VeryFast, however, treated the problem in a new way allowing the computation time for image resizing to be transferred from test to training time. The idea is partly inspired by the work of Piotr Dollár, suggesting to only scale the input images N/K times [72], K simply being a down-sampling factor of ~ 10. VeryFast leveraged this idea and essentially reversed it, sticking with the Viola-Jones principle that the image features should be scaled and not the image. The concept of feature scaling is best illustrated in Figure 2.15.



(a) Naive approach for dealing with feature scales.

(b) N/K Scales for dealing with feature scales. This is an *approximation* across scales.

Figure 2.15: The images shown in (a) and (b) above highlight the difference in approach taken by Benenson et al. to deal with multi-scale detections. Image (a) generates feature models for every scale and is hindered by this. Image (b) uses only the input image at test time and approximates features across multiple scales with negligible accuracy penalties. Image sourced from Benenson et al. work [8]

.

This concludes a fairly comprehensive review of the key contributions towards object detection over the last decade, at least for sliding window approaches. With the very recent advent and widespread adoption of deep neural networks though, the game has simply changed. Features were no longer constrained by poor human design and suitable optimisation methods were developed to take advantage of increasingly large labelled datasets. One of the emergent applications of this was improved object detection methodologies, which we can now evaluate.

2.5.4 Convolutional Neural Network Detection

Machine learning has emerged to be arguably the biggest influence in computer vision over approximately the last five years. The step change was noticed after the work of Krizhevsky [41] made huge performance gains in the ImageNet large scale object recognition challenge, prompting leaders in the field to take notice. We will examine his work and other related art later in the thesis, when we develop our own recognition algorithms. For now it will suffice to describe the general nature of his proposed solution.

We have discussed neural networks and stated their impressive learning ability for extracting inferences from data. That statement, while true, ignores some aspects about the scalability of NNs that become clear when applying to large images. Given each pixel of an image can essentially be used as an input feature, where colour images have 3 planes worth of pixels and serve as input to *each* neuron, the parameter space can quickly explode for reasonably sized images. The standard feed-forward ANN architecture also ignores the locality and nature of pixels in an image structure. Thus, a special architecture specific to neural networks was developed to take advantage of the 2-dimensional nature of images. This special case is the Convolutional Neural Network and it has become the workhorse of machine learning tools for vision applications.

In general a CNN works by sliding or *convolving* a convolution kernel across an input image to create an activation map, also known as a feature response map. At each hidden layer of a CNN there can be many of these kernels and subsequent feature maps, which are then downsampled by a pooling operation and a non-linearity is introduced. As images propagate through more layers of a CNN it effectively learns higher levels of abstraction for each object class, using error minimisation to tune the kernel weights. The final layers of a CNN tend to be standard neural network layers, where the spatial component of the images is lost and it becomes a straightforward classification problem using the extracted high level features. Once a CNN has been trained and the error is sufficiently low, the static kernel weights represent low-level object features that have been automatically discovered through the supervised learning scheme. It is this automatic feature discovery by the machine that makes CNNs very powerful tools for image recognition and other applications, as it removes the human design element from the task [41].

Bearing this brief summary of how a CNN works in mind, we now have the capability to obtain excellent object classification results. But we still require methods to *locate* objects of interest in an image, which is where adapted CNN methods enter the fray. One of the most notable recent detection methods combines region proposals with CNNs, which is aptly named *R-CNN* [79]. In this landmark work Girshick et al. acknowledge the plateauing progress of object detection research, which had become more reliant on complex ensemble methods. They address the detection problem using CNNs to compute features for thousands of bottom-up region proposals extracted from an input image, which are then classified as the correct object or rejected. It is incredibly simple and exploits the high capacity offered by CNNs along with a supervised learning trick to fine tune a *pre-trained* network on scarce data, which provided a significant accuracy boost. At the time of publication this

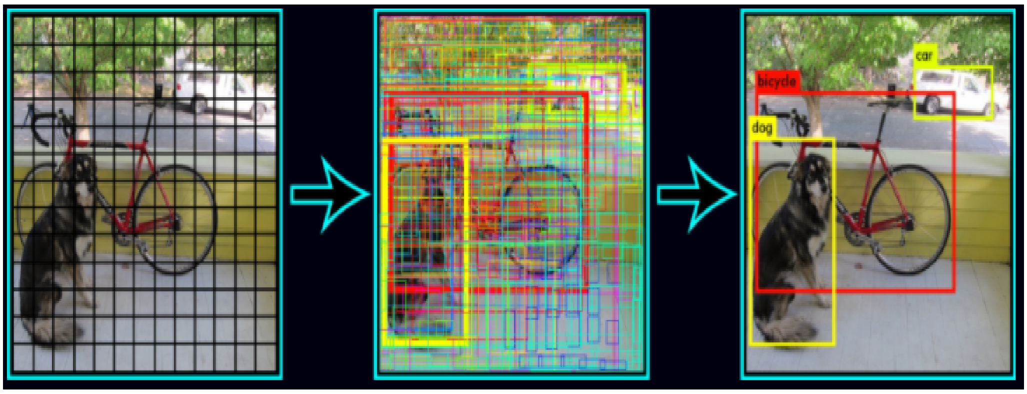


Figure 2.16: Pipeline of YOLO model is presented. A single trained neural network is applied to the full image above, showing a grid overlaying a park based scene. The trained network divides boxes into regions and generates associated probabilities for these, which are then weighted according to the probabilistic values. The end result in this case correctly identifies and localises a dog, bicycle and car.

method improved upon the next best algorithm by roughly 30% and proved very elegant to develop and deploy. It has since been advanced to its current state, the Fast R-CNN [80, 81]. The obvious improvement is the speed of execution compared with the initial solution.

Lastly, the most recent advancement made in this area is the aptly named You Only Look Once (YOLO) and succeeding YOLO9000 model [82, 83]. The YOLO model addresses the fact that traditional detection approaches usually rely on repurposed classifiers. This work re-frames object detection as a regression type issue for many boxes seperated spatially with associated class probabilities, which is highlighted in Figure 2.16.

This allows YOLO to be a single trained neural network optimisable for detection tasks. The processing time can range from around 45 frames per second at maximum accuracy up to 155 frames per second for less accurate output, but still double the precision of competing real-time detectors. The YOLO9000 model offers slight improvements and can recognise significantly more object categories, approximately 9000 in total.

This short review of where the state-of-the-art lies in terms of object detection is only a glimpse at the current research field, where we very briefly covered one of the most notable object detection methods of late. It incorporates modern machine learning methods and exhibits excellent performance towards detection tasks in imagery. Convolutional Neural Networks will be given a full examination in the later chapters of this thesis.

However, while increasingly accurate detections and classifications can be obtained, all algorithms, including VeryFast and R-CNN, suffer from occlusions, clutter and dynamic illumination. There is much progress to be made if the dream of complete machine comprehension and interaction is to be realised. Although feature selection and generation has been discussed above as an area for improvement, it is not the whole story. A human visual system can track individuals in a crowd or deal with heavy occlusions because of prior, learned experience. If this mechanism of higher level understanding could be given to a machine, it would surely benefit detection accuracies and enhance overall capability. This theme can be encapsulated as the utilisation of context and is explored in the relevant section in this chapter, following a brief discussion of occlusions and thermal detection algorithms.

2.5.5 Tracking Through Clutter & Occlusions

Occlusions can generally be defined in three separate ways. The first is localised in space and is referred to as a partial occlusion, an example being a pedestrians lower half being obscured by a bin or suitcase etc. The second is localised in time and is referred to as a full occlusion. A good example would be a jogger being fully obscured as they run past a row of trees. The last type is scattered occlusion and this differs from the previous types in that is not localised in either space or time. Snow, heavy rain, fog and thick foliage are all examples of scattered occlusions. Occlusions are troublesome for detection systems in that they either throw initial detections, since features do not match half-pedestrians, or they degrade most tracking algorithms since the object will go out of view and become unrecoverable. Whilst target objects will still be in view for scattered occlusions they still make detection and tracking difficult as no correlation between nearby pixels can be assumed [84].

Work related to detection and tracking in occlusions is quite broad but yet the problem still persists in the majority of detection systems. Some example tracking methods that are well known include the Kalman & extended Kalman filter [85] and the particle filter or the boosted particle filter [86]. The Kalman and particle filters work by predicting where target objects are going to be in the next frame according to some learned model. This is useful for short lived occlusions but it fails when occlusions last longer. The boosted particle filter combines this prediction tracking approach with a detector to recover the original object tracking after a long occlusion. Finally, the scatter tracker from [84] deals with non-localised occlusions by using a similarity metric in conjunction with a spatial prior.

Clutter is harder to define than occlusions. Generally, clutter is the background to foreground objects that make it hard to determine what the said foreground is. In other words, it *increases* ambiguity within a scene. A good way to envision this problem is to imagine a scenario of placing a post-it note on a desk for a work colleague or friend. If this desk is clear, or uncluttered, the post-it will be easily visible and hopefully noticed as intended. However, imagine the desk is now cluttered with other post-it notes, papers and books. It becomes much more difficult to place the post-it with hopes of it being successfully noticed. This problem encapsulates why clutter is troublesome for computer vision systems, as it can be extremely difficult to identify the foreground from the background in cluttered environments, i.e. the correct post-it placed on a desk of post-it notes.

2.5.6 Thermal Detection Algorithms

Thermal or IR detection is usually applied in surveillance or military scenarios, especially when night-time capability is required. The thermal band is a much less rich source of information than RGB and as such there is less literature on thermal detection algorithms. The majority of existing methods follow the general idea of localising hot-spots or shape matching within an image via thresholding and possibly motion between frames [87, 88].

Recent work includes a shape context descriptor with and Adaboost cascade classifier, which is shape matching with machine learning principles, while a template based method also exists from Davis & Keck [89, 90]. This template method is carried out in two stages where the first step is to create a contour saliency map of the target object within thermal images, which in their case the target is human in appearance. These saliency maps are averaged and form a screen that can be used in a multi-resolution screening step, returning potential target objects. This method in particular exploits the invariance of edge information in thermal images.

Very recently a significant effort to advance this area has been driven by Berg et al. from the computer vision group at Linköpings universitet, especially in the area of object detection and tracking for thermal imagery. These advances are made in two key areas, the first being a labelled thermal dataset and evaluation protocol is publicly available [15, 91]. Secondly, Berg et al. offer a robust short term, single object tracker adapted for thermal infrared imagery using well known recent template based tracking methodologies [92]. The developed object tracking method is

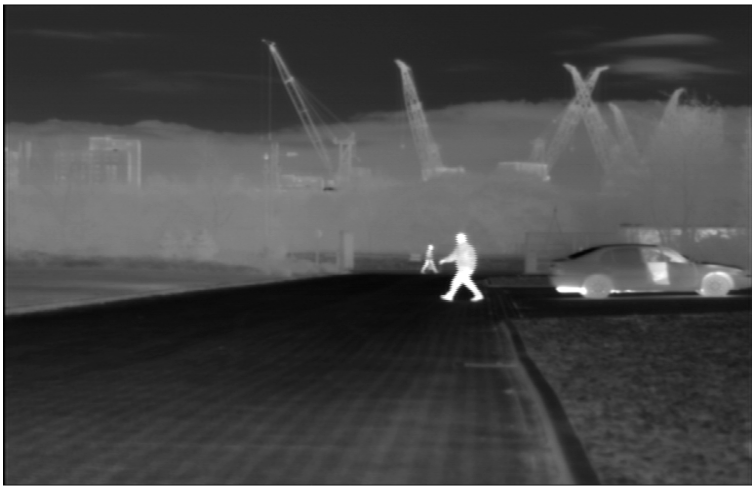


Figure 2.17: Image captured by a passive thermal sensor. This is a typical example for the modality and clearly illustrates thermal hotspots for pedestrians and cars. Image collected at Thales Glasgow site using a Catherine MP camera .

suited to thermal imagery where channel coded distribution fields show a distinct advantage over spatially structured features.

Overall, the potential for using the IR modality to enhance a vision system is clear. It can tackle the problem of dynamic illumination present in RGB systems and is highly useful for target detection through scattered occlusions, given it can pierce fog and foliage better than colour band sensors [1]. A representative image is presented in Figure 2.17. Some drawbacks for this modality are the lack of texture information in images as well as the lack of contrast between the foreground and background if the temperatures are the same. Both RGB and thermal imaging seem to have ever-present problems. One identified area that may offer improvements and solutions to some of these issues is the role of context for computer vision.

2.6 Utilising Contextual Information

Thus far, SOA methods have been explored and the relevant inadequacies identified. This section will discuss how it is possible to achieve higher level image processing techniques by discussing the use of contextual information and psychophysical studies to gain an insight into the human visual system. A good example to illustrate this point is presented in work by Harding and Robertson [93], where it is confirmed humans subconsciously search for salient regions of an image, regardless of tasks. A similar point is made by Hanson and Essock where they demonstrate the tendency for humans to search images horizontally for people, simply because it *makes sense* to do so. Pedestrians are most likely not going to be traversing a sky region in an image and this reasonable knowledge is exploited by the brain automatically when searching an image [94]. Hence, we cannot simply treat every pixel within an image as equal when it is not the case. The evidence in the literature points towards this eventuality and in order to give machines a human-like capability for detection tasks, extra information must be given to the machine. This extra information is analogous to life experience gained by humans, aiding how we perceive world scenes. This is known as contextual information. It must be acknowledged scene context is slightly different from visual saliency, but both highlight the point that pixels are not best treated equally.

Before exploring how context can offer improvements to scene segmentation and recognition systems we must first identify a clear definition for context. Somewhat surprisingly the literature proves sparse for definitions that were not wholly vague. One example definition is offered from Strat, who states context as *"any and all information that may influence the way a scene and the objects within it are perceived"* [95],. In a sense this is correct but it does not help identify sources of contextual information, where *"any and all"* could be a huge number of things. Some wide ranging sources of context are provided by Divvala et al [18] and presented in Table 2.2.

From these generalised definitions it appears context is essentially extra information that potentially improves overall understanding of a scene. Although the list presented in Table 2.2 is not exhaustive the work contained in this thesis only exploits a handful of these sources. The three key context based ideas utilised in the data chapters are *Temporal*, *Semantic* and *Local Pixel* context. Temporal context is loosely enforced in Chapters 3 and 4 to incorporate information from previous video frames. Semantic context is employed in all chapters, either through region or object presence and local pixel context is predominantly featured in Chapter 6 to aid long range target recognition.

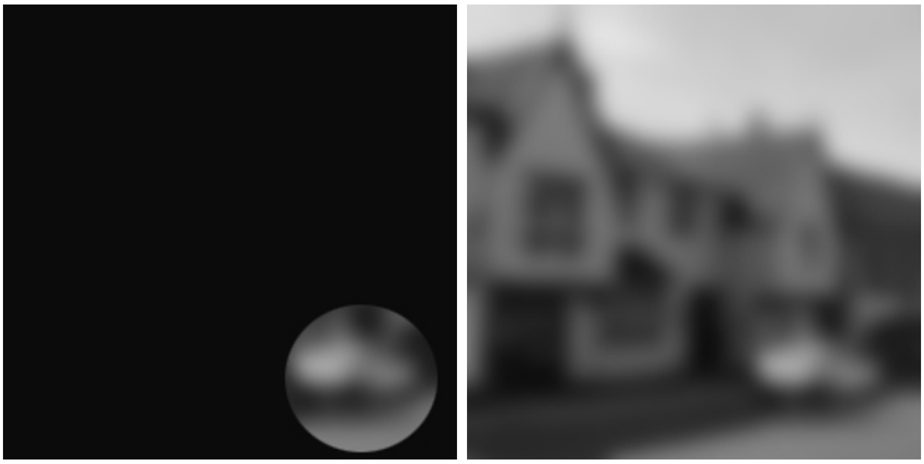
Humans possess this underlying comprehension of real world events due to past experience. Machines, on the other hand, do not possess this prior knowledge and would benefit entirely if given this extra understanding. Whether this additional context comes from new sensor information altogether, such as inputting GPS location data as geographic context, or is produced from manipulating original pixel level data into a different representation, the end goal is always to achieve a higher level scene understanding. An example of how humans can use context to interpret a scene is presented in Figure 2.18.

Sources of Contextual Information	
Context Source	Description
Local Pixel	Window surround, image neighbourhoods, object boundary/shape
2D Scene Gist	Global image statistics.
3D Geometric	3D scene layout, support surface, surface orientations, occlusions, contact points, etc.
Semantic Context	Event/activity depicted, scene category, objects present in the scene and their spatial extents, keywords.
Photogrammetric	Camera height, orientation, focal length, lens distortion, radiometric response function.
Illumination	Sun direction, sky colour, cloud cover, shadow contrast, etc.
Weather	Current/recent precipitation, wind speed/direction, temperature, season, etc.
Geographic	GPS location, terrain type, land use category, elevation, population density, etc.
Temporal	Nearby frames (if video), temporally proximal images, videos of similar scenes, time of capture.
Cultural	Photographer bias, dataset selection bias, visual clichés etc.

Table 2.2: The table above provides a variety of sources for contextual information. *Any and all* may be used to garner a better scene understanding [18].

The effective use of context is closely tied to how the human visual system performs in detection and recognition tasks, simply because if we want to allow machines the same ability we must first understand the mechanics behind our own system. This notion leads us to some famous psychophysical studies providing crucial insights into how we use visual information for scene understanding. The vast work of Biederman offers many insights into how humans quickly interpret scenes using semantic relationships between objects [96–98].

Biederman also posits five classes of relationships between objects and its setting. These are essentially a set of rules generalising how objects should be organised within a scene to make sense. These relations are *1.* Interposition - objects interrupt their background. *2.* Support - objects tend to rest on surfaces. *3.* Probability - objects tend to be found in some scenes and not in others. *4.* Position - given an object is probable in a scene, it is often found in some scenes but not in others. *5.* Familiar size - objects should have a limited set of size relations with other objects. These definitions can be found in [97] from Biederman et al, where the study used these definitions to prove humans utilise them to interpret a scene when presented with an image snapshot lasting only $\approx 150ms$. Studies like these show there is a much



(a) Blurry snapshot of an image. Is the object or scene identifiable with just this snapshot?

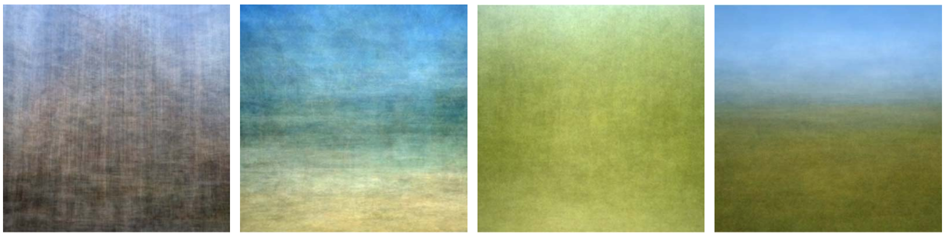
(b) Complete blurry image of the same scene. The scene and object within the viewfinder should now be instantly recognisable.

Figure 2.18: The above images should be a clear illustration of how humans use context to infer object classes and scene understanding. Even with extensive image degradation, such as blurring above, it is easy to recognise the object and environment when given the whole picture. This indicates that we are quite influenced by contextual information and giving the capability to a machine should be fully explored. Image taken from [9].

deeper process at work behind the human visual system. One such attempt to create a computational scene recognition model is developed by Oliva and Torralba [10].

The work by Oliva and Torralba takes the original low-level pixel data and creates a new, global representation of a scene that can be interpreted at a glance without the need for segmentation or region processing This representation is named the *spatial envelope* and may also be thought of as similar to scene *gist*, which is also an abstract scene representation proposed by Friedman [99]. The model uses spatial and spectral information to show specific information about object shape or identity is not an absolute requirement to categorise a scene overall. Scene categories include streets, forest, coastline, buildings etc. Some example representation images are provided in Figure 2.19.

The holistic representation of a scene category successfully illustrates it is possible to construct a scene as a single quantity that does not require any object detection or recognition information. This idea is used in a follow up piece of work by Torralba, where the goal is to use a similar scheme for incorporating contextual information in object representations and use it for object detection purposes. The method is known as contextual priming for object detection and relies on using



(a) Tall Buildings (b) Coast (c) Forest (d) Open Country

Figure 2.19: The four images above are generated using the spatial envelope from Oliva and Torralba. The scene categories are given for each image and they represent a holistic representation of that category. Although it can appear a vague way to represent a category, it can inform the viewer of a semantic category with a glance, without the need for specific shape information. Image source [10]

statistics to model low-level pixel data along with object-centric data, such as size and location. The end result is a context aware scheme capable of selecting task driven regions (focus of attention) in an image as well as automatically inferring image scales. This is achieved via probability statistics and Bayesian mathematics [9].

Contextually priming for object detection is beneficial as it can be used to manage detection algorithm workload, directing resources to a *primed* search area. Furthermore, this method also does not fail because the background produces false positives or distractions, it exploits this information to enhance our knowledge of the scene and object properties. This should be useful when dealing with cluttered images.

Context can also be seen to offer improvements to tracking tasks. Recent work by Borji et al. create an adaptive tracking scheme by using a learned model for background context [100]. It essentially adapts the object descriptors as and when the scene background undergoes strong changes. For instance dynamic illumination could be a troublesome issue and present a rapidly changing background. The method uses particle filters (a weighted set of points on an object that are updated each time step and predict object movement) and certain contrasting colour based components. The end result is a tracker capable of dealing with occlusions and it is shown to pick up the same object after full occlusions. One of the drawbacks of this method is that a user interaction stage is required for every new video sequence in a learning phase. This hints at it not being deployable for real-world tasks, where hard real-time processing is essential. Another example of a context based tracker can be seen in the work of Maggio and Cavallaro. This method learns scene context with a Bayesian, probabilistic approach and handles the occlusion scenario in a rather novel way by modelling target *births* after occlusion events as well as modelling the spatial layout of clutter. The end result is a multi-object tracker improved by using scene context [101].

2.7 Conclusion

A system capable of utilising both thermal and visible modalities, whilst incorporating scene context to vastly improve performance is a highly desirable prospect. Improving situational awareness in military or surveillance tasks is always beneficial as human processing is limited. This is why we strive to build target detection systems capable of flagging potential threats, with the hope it is much quicker than our brain. This is because it is likely a wrong decision is made in a potentially dangerous situation. Hence the need for tools to help people in these situations make better and more informed decisions. Furthermore, it is also desirable to have fully automated and robust surveillance because it is infeasible for humans to solely perform this task on a massive scale. It is proposed that utilising machine learning methods and context can aid segmentation and target recognition performance, which ultimately will be a step towards realising some of these overarching goals.

Sensors are used to collect information and each sensor will have limitations. Visible sensors are weak in changing illumination which complicates object detection algorithms. Whilst they provide a rich source of information they will become effectively useless at night-time. Thermal sensors essentially address this issue and also have the benefit of identifying hot targets through scattered occlusions. To combine both sensors provides an obvious advantage. These sensors can then be used to collect pixel information and a scene can be constructed. An overview of segmentation algorithms was presented, with great detail given on SLIC and a graph based method. Both tackle a subjective segmentation problem but fail to perform the task as well as required. Furthermore, object detection practices were reviewed from the classical treatment to current deep learning methods. Sources of context and scene modelling were also discussed, highlighting the many benefits to be gained from incorporating contextual information. The most important aspect covered, however, is the treatment of foreground and background region segmentation, especially in current state-of-the-art semantic segmentation methodologies.

Presently machine learning is driving semantic segmentation, where foreground objects are parsed from surrounding background regions and classified. This makes sense on one level as foreground objects are typically what we are interested in and are the defining feature of a target scene. Yet, if we extrapolate current methods to our problem space in the defence domain, where targets can exist at long ranges in challenging rural environments, they will fail. This is because targets have very small signatures and cannot be easily parsed or identified. This is especially true in low light scenarios as we have already demonstrated. In this scenario it is the surrounding regions and scene context that become the defining feature of a scene, which we aim to utilise in order to achieve our goal of robust autonomous target recognition and enhanced situational awareness.

The key ideas going forward address this gap in the literature, exploiting the tightly coupled relationship between foreground objects and surrounding background regions. Having adequately explored the project motivations and placed it in the context of current literature, we can now begin to discuss efforts contained within the thesis towards achieving our overall goal. The next chapter examines the relationship between foreground and background regions in cluttered environments.

Chapter 3
Recovering Background Regions in Cluttered Scenes

To address the gap identified in related work, with regards to foreground and background segmentation, we begin by focusing on developing a suitable mechanism to extract underlying background regions in cluttered urban scenes from colour video sequences.

An existing superpixel representation is sufficiently modified to allow region merging, predicated by a defined similarity metric. We present a method that aims to not only obtain stable regions but is capable of leveraging emerging foreground context to recover underlying background segments. Prior temporal and spatial information are also explored to see if they are beneficial to the segmentation process.

These are the key considerations for this chapter and they are fully discussed within. By exploring these ideas we discover the presence of a foreground object indicates something about the scene structure as a whole, which we exploit to our advantage. The findings are summarised in Section 3.5, the conclusion of this chapter. This work was presented at the IEEE International Conference on Image Processing, 2015 [23].

3.1 Introduction

Segmentation and detection problems are often treated independently where methods that identify and address the coupling are less in evidence. We are motivated to explore the relationship between image segmentation and object detection as it is directly related to foreground and background regions of an image, which is a key aspect of the overall project goals. The grand aim of the work presented in this

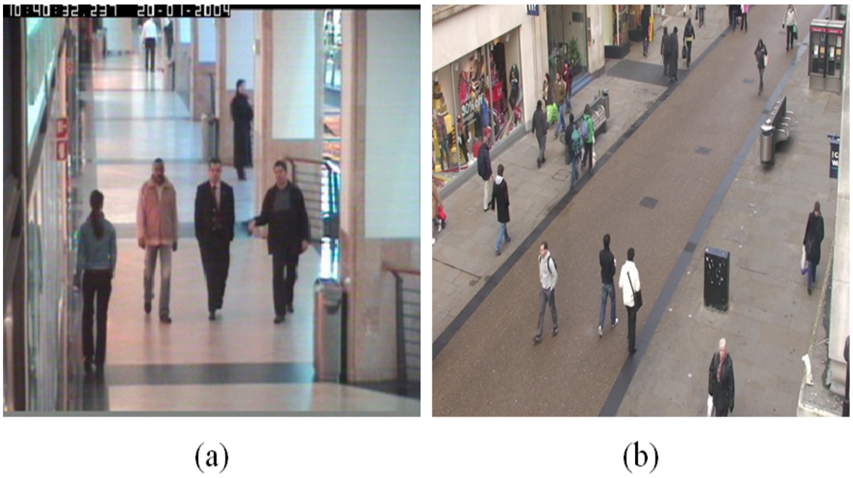


Figure 3.1: Image (a) shows a crowd of people walking through a shopping mall from a static surveillance camera viewpoint [11]. Image (b) presents a crowded Oxford street scene [12]. Both examples typify a standard surveillance scene.

chapter is to develop enhanced scene understanding by exploiting the *foreground context* given by an objects presence within a scene. In order to achieve this the background needs to be segmented into regions which may correspond to regions of activity. A typical urban scene containing pedestrians is shown in Figure 3.1. In such environments it is not normally possible to intervene in order to image an uncluttered scene, where obtaining regions would be trivial. It is intuitive that, over time, the presence of foreground objects tells us something important about the background .

We propose observing foreground tracks throughout an urban scene is a positive indication for closely related regions, suggesting they are likely candidates to be merged in a segmentation process. This would result in segments more closely resembling the true background regions underlying frequent foreground activity. We therefore develop a new segmentation method enabling this influence to be computed. Additional incorporation of spatial and temporal priors allow successful integration of all available video information. We test the method on pedestrian videos only, although our algorithm would apply equally well to other structured activity scenarios such as vehicular traffic.

As motivation, consider an everyday scene that is generally well-populated with people, such as a train station, airport or shopping mall. These scenes will typically have few distinct regions which are quite similar in appearance. For instance an airport lounge will likely have a main concourse for people to occupy, some wall/shop fronts and perhaps windows or pillars. The application of a stand-alone segmentation algorithm would fail to recover key background regions for such a scene. This is because foreground objects, i.e. people, are acting as clutter on the background regions. They will have distinct distributions when compared with background regions, regardless of what features the segmentation algorithm exploits, e.g. colour or texture information. In essence the background and relative structure of a cluttered scene can be potentially unrecoverable when treated independently with a standard segmentation approach.

The work developed in this chapter provides a method for estimating true background regions in scenes exhibiting dense foreground object clutter. Algorithms utilising the Bhattacharyya coefficient as a similarity metric towards merging regions already exist [102, 103]. We extend the best of these and the initial region merging process is outlined in section 3.3.1.

The focus of this chapter the utilisation of foreground context to improve scene understanding, especially in crowded urban imagery. In doing so we advance the contextual framework described by Letham et al. to smooth the final segmentation output using temporal and spatial priors [104]. This Bayesian framework is altered to suit our problem formulation to overcome the reliance on pre-fabricated ground truth by adopting a purely data-driven approach. An extensive evaluation using standard segmentation metrics is provided for full label maps and dual layer outputs, using the CAVIAR [1] and Oxford datasets [12].

3.2 Related Work

Background (BG) subtraction techniques present a similar vein of research, especially when such methods address crowded scenes in surveillance scenarios [105]. Although the output from BG subtraction differs from our desired goal, the challenge of dealing with foreground clutter remains. We address this by exploiting object information to enhance BG segmentation. Attempts to integrate segmentation and detection in the literature are relatively scarce. Gould relies on using segmented regions to aid object detection and classification processes [106], while Wojek's approach creates a single CRF to jointly model both processes [107]. Gu has also shown segmentation to be an effective aid in object classification tasks [108]. Generally it is deemed desirable to segment out and identify foreground objects. It is also suggested to avoid segmenting out many instances of an object class, such as a row

[1] http://groups.inf.ed.ac.uk/vision/CAVIAR

of cars, as a whole region [106]. However, we are purposely attempting to segment only FGBG regions to avoid object labelling associated with semantic segmentation.

For semantic segmentation, where object classes are simultaneously detected and segmented, using a CRF framework has proven very successful [109]. Foreground (FG) object information is identified for incorporation with pixel labelling, via a joint reasoning scheme. Additionally, recent work has used distinctions between Foreground-Background (FGBG) regions in video streams to effectively segregate each layer [110, 111].

Similar Approaches: The bilayer approach explored in [110] and the contextual framework of Letham et al. presented in [104] are likely the most similar works in relation to this data chapter. Given the developed method naturally incorporates key elements of [104] this should come as no surprise. The connection to that of Sun et al. [110] is less explicit. Mainly it is the concept of dividing an image into a two layer structure of interconnected Foreground and Background regions that is similar. Sun et al. posit that previous methods treat this problem poorly and can be solved using a bilayer approach, a notion shared in our work. However, the key difference between [110] and ours lies in what each algorithm is trying to *extract* from the image or video footage. Sun et al. are focused on using the BG layer to inform the segmentation of FG objects, while we are trying to solve the opposite problem.

In other words, FG knowledge can improve the extraction of BG regions and global scene comprehension as a whole. The algorithms will aid the development of a robust region segmentation process to enhance overall scene comprehension, feeding into our larger goal of improved autonomous target recognition, whereas

3.3 Background Segmentation Method

The basic outline of the algorithm is presented in Figure 3.2. In summary we take an initial segmentation of an image and then use FG knowledge to influence future segmentation refinements. The region merging process follows directly from an initial segmentation stage which is obtained from a superpixel method - the Simple Linear Iterative Clustering (SLIC) algorithm [3]. A very brief overview of SLIC is provided in section 3.3.1, given we have explored the technical details in full in Chapter 2. The proposed region merging technique and subsequent processes are introduced in

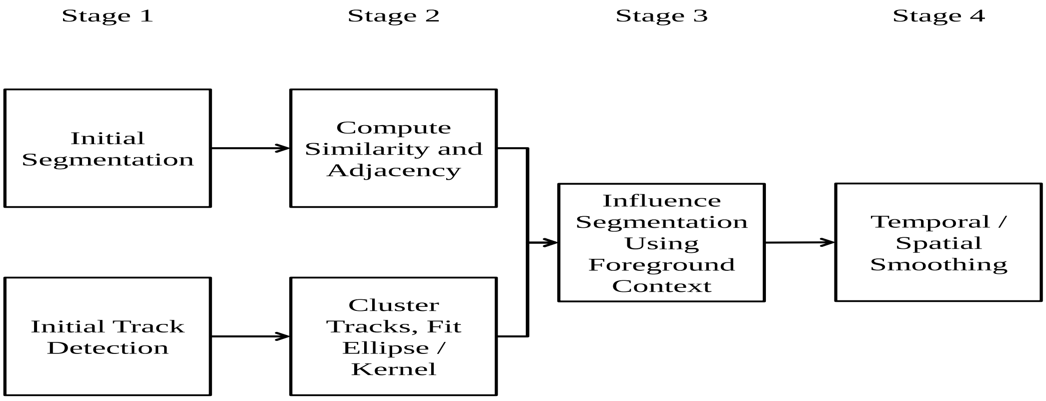

Stage 1
Stage 2
Stage 3
Stage 4
Initial Segmentation
Compute Similarity and Adjacency
Initial Track Detection
Cluster Tracks, Fit Ellipse / Kernel
Influence Segmentation Using Foreground Context
Temporal / Spatial Smoothing

Figure 3.2: Schematic of our method showing main steps. Initial tracking and segmentation processes happen simultaneously and are shown as parallel flows at Stage 1. Stage 2 involves the collation of observed foreground tracks, while superpixel similarity is computed. stages 3 and 4 combine this information to extract underlying background regions.

following sections.

3.3.1 Region Merging

Superpixels are groups of pixels deemed to be similar in nature, usually through colour or proximity information [60]. Such methods essentially present a form of curtailed region segmentation by avoiding over and under-segmentation, yet adhering very well to image boundaries [112]. Thus we propose superpixels to capture an underlying image structure and if merged appropriately, form whole contiguous regions corresponding to this image. The input for our region merging process is a superpixel label map via SLIC. SLIC works by designating each image pixel a 5-dimensional feature vector allowing a normalised distance between pixels to be calculated. Then k-means clustering is performed to group pixels and a clean-up step is implemented to enforce connectivity of superpixels whilst ensuring there are no stray pixels. The output of this SLIC algorithm serves as our initial input segmentation.

To appropriately merge superpixels we compute a similarity metric to determine how *close* each superpixel is relative to other superpixels. The Bhattacharyya distance, B_{dist}, effectively provides a probability of superpixels being dissimilar given colour channel information [113, 114]. The B_{dist} operates on normalised histograms for each colour channel, which is possible as normalised histograms are discrete estimates of probability density functions (PDFs). For each superpixel (SP) we can then compute a normalised histogram H using all pixels contained within each superpixel, for each information channel c, corresponding to the colourspace used. This normalised histogram for each colour channel is represented as H_C. In our experiments we use the conventional RGB-space, so the dissimilarity or *distance* D_H between two superpixel distributions SP_i and SP_j is given by Equation 3.1.

$$D_H(SP_i, SP_j, C_{RGB}) = B_{dist}(H_C(SP_i), H_C(SP_j)) \tag{3.1}$$

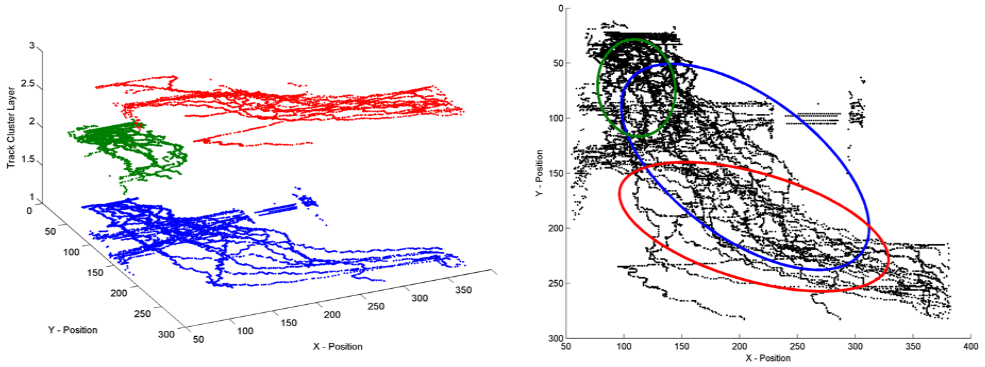
The distance D_H is computed for each colour channel and multiplied through. We can then create a dissimilarity matrix, BD, populated by D_H for each SP compared with all other SP distributions. This array will have a zero entry diagonal and be symmetric. Once BD is obtained, a global threshold T_g to control the merging process can be determined. If we let J denote the indexes of the columns of matrix BD, W_t as a weighting factor and ij are the usual matrix elements, then the global threshold T_g can be computed as shown in Equation 3.2.

$$T_g = mean(\underset{j \in J}{min}(BD_{ij})) \times W_t \tag{3.2}$$

Essentially we are taking the average of the minimum for each matrix row and the weighting factor allows a certain degree of regulation, to either relax or limit the threshold if desired. An adjacency matrix can then be computed using 8-neighbourhood connectivity to determine how SP label regions are related spatially. In other words we determine which regions are located next to each other. Combining this information with BD allows merging between SPs to occur according to the conditions outlined in Equation 3.3.

$$M(SP_i, SP_j) = \begin{cases} 1, & \text{if } BD_{ij} < T_g \text{ and } SP_{ij} \text{ adjacent} \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

The merging function M referred to in Equation 3.3 only allows merging between adjacent superpixels if they are lower than T_g. If merging is desired a list of all candidate superpixels fulfilling the joining conditions, then it is performed in one cut with the output label map renumbered from 1 to n regions. However, if we wish to influence the segmentation process via FG context we do not want to merge superpixels immediately. Instead we only require the dissimilarity array BD for Stage 3, as shown in Figure 3.2.



(a) (b)

Figure 3.3: Image (a) shows a 3D view of track clusters sampled from the CAVIAR ground truth pedestrian tracks, highlighting distinct clusters of activity. Image (b) presents a top-down view of ellipses fitted to Caviar track cluster samples, track points are shown in black. Three ellipses are generated, corresponding to one per layer of track clusters shown in the left hand image.

3.3.2 Exploiting Foreground Context

Before we utilise any FG context it has to be observed. This is shown to be happening in parallel with initial segmentation processing prior to Stage 3 in Figure 3.2. Given the problem posed of segmenting background reliably in cluttered scenes, there is a clear assumption that the imagery in question is well populated by FG objects. To collect FG track data it should simply be a matter of implementing an object tracker. That is not the primary focus of this work and the benchmarked datasets already have pedestrian ground truth, which we use in our experiments for the Caviar datasets. Real tracking information, obtained from Baxter et al. is then utilised for our experiments on the Oxford datasets [115].

Individual tracks are processed using a recent trajectory clustering technique, based on mean-shift, in order to determine an underlying structure in the FG data and remove spurious information [116]. An example of the track clustering is presented in Figure 3.3. An ellipse fitting method can be applied to each layer of track information, also illustrated in Figure 3.3 [117]. Masks directly used to influence segmentation, via altering the dissimilarity scores in array BD, are derived from FG data. We assume any ellipses fitted to track clusters will have the least associated error nearer the centre of the ellipse than towards the edges. To reflect this we fit a kernel to each ellipse mask, weighting the FG influence accordingly using the kernel

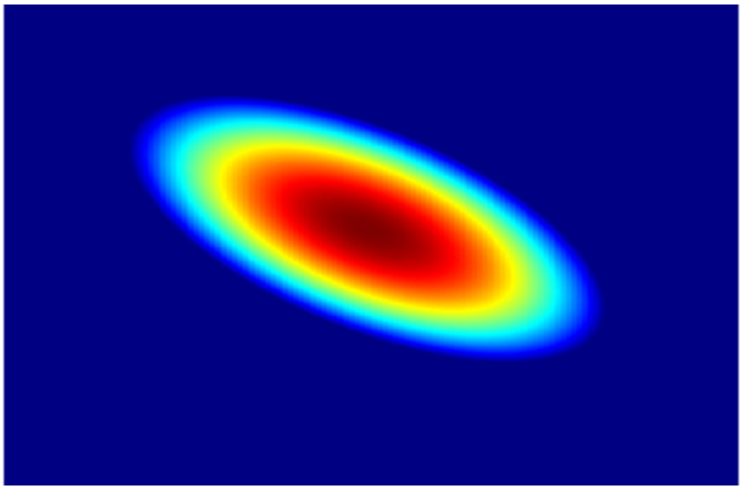


Figure 3.4: The image is a simple colourmap showing the relative weights of the Epanechnikov kernel after convolution with a track cluster ellipsoid shape, generated via the earlier FG observations. The goal of this kernel is to positively affect the similarity scores for merging superpixels, based on the FG evidence. In the image, blue relates to no change while red corresponds with a large alteration to the underlying BD scores.

weights K_w.

To actually alter the scores in array BD we overlay each mask onto the image plane of the initial segmentation label map. The kernel value corresponding to each SP centre pixel location is then retrieved, which is carried forward and multiplied by the SPs dissimilarity score *relative to adjacent SPs* using the relationship $1 - K_w$. The Epanechnikov kernel [118] is employed to achieve this and is given by the kernel function described in Equation 3.4, where $|u| \leq 1$. This now influences the region merging process using FG information and a visualisation of the Epanechnikov kernel weights, after convolving with one of the fitted ellipses shown in Figure 3.3, is provided in Figure 3.4. We can then calculate a threshold and merge in the same manner as described in section 3.3.1.

$$K_w = \frac{3}{4}(1 - u^2) \tag{3.4}$$

The desired output is a segmentation with large, contiguous regions due to the effect of FG context. By combining the elliptical masks into one large mask, we overlay it on the output segmentation and determine the largest region or regions within the ellipses. These are deemed to be Background-Foreground (BGFG) regions given they lie in ellipses fitted from FG data, effectively being the backdrop to FG objects. Having identified regions as BGFG, it is then possible to binarise the output by considering every other region as one which FG objects do not occupy. This can be denoted as the Total Background layer (TBG) . The binary segmentation is used to help smoothing as explored in section 3.3.3. Once this step has been completed, a final refinement stage can be implemented and is shown as Stage 4 in Figure 3.2.

3.3.3 Temporal & Spatial Smoothing

We adapt a Bayesian smoothing framework defined in prior work to introduce temporal and spatial information for each segmentation of a scene frame [104]. The original method is developed for binary classification of multiple region types, using prefabricated ground truth information to provide the spatial priors for each region type. In this chapter we only want to perform smoothing by considering a binary region approach, namely the BGFG and TBG layers produced from the method outlined in Sub-section 3.3.2. Bayes theorem is used to compute the probability for either of the binary regions to exist, provided in Equation 3.5.

$$P(R|D) = \frac{P(D|R)P(R)}{P(D)} \tag{3.5}$$

The terms in Equation 3.5 can be described in the following ways. The probability of a region R to exist, given the detection D of said region is the posterior and can be shown as $P(R|D)$. The true positive rate when comparing the BGFG region to the combinatory ellipse mask is $P(D|R)$. Finally, the temporal and spatial prior is given by $P(R)$ and $P(D)$ is simply a normalising constant.

The prior can be explained as a weighted summation of the prior probability dependent on the current FG track knowledge, $P(R|T)$, and the posterior probability of the previous frame $P_{k-1}(R|D)$. Essentially the term $P(R|T)$ is a spatial prior based on the ellipse masks derived from FG track fitting. thus, the term $P(R)$ can be calculated using Equation 3.6. Referring to this equation, w lies in the range [01] and controls the smoothing contribution from the temporal or spatial term. Once $P(R|D)$ is computed, we again alter the scores in the dissimilarity array BD, corresponding to which binary layer (BGFG or TBG) the remaining SPs lie in, using the obtained $P(R|D)$ value.

$$P(R) = (1 - w)P_{k-1}(R|D) + wP(R|T) \tag{3.6}$$

3.4 Experiments

We evaluate this method on the benchmarked data using standard segmentation metrics. A random subset of ground truth tracks is sampled from the CAVIAR data and clustered using mean-shift based multi-feature trajectory clustering [116]. This initial experiment allows us to vary key parameters to discover their effect and guide future use. We then proceed to test on the Oxford set using empirically optimal parameters and on-line track information collected over the first 20 seconds of video footage. The region extraction process was then implemented on 300 subsequent frames using 50 superpixels and a strict threshold. We finally compare our results to the EGB segmentation method [2], Normalized Cuts (NCut) [119] and variations of our own algorithm. For completeness, a naive median filtering preprocessing step is also utilised with EGB and NCut in an attempt to remove foreground objects from the scene using the simplest conceivable method. However this method degrades significantly for more cluttered scenes.

Let us now discuss the parameters used in the experiments. For the initial segmentation and proof of concept we use SLIC with input 50, 100 and 150 superpixels, with the compactness factor m fixed at 20. The global threshold weighting factor W_T was set at 0.5 and 1, strict and relaxed respectively. Five weights of smoothing were chosen, where w ranged from 0 to 1 in 0.25 increments. These were used to find the best performing setting for the proposed method. All results presented are generated using 50-SPs with a strict merging threshold, where we take a median average of metric output from the different context weightings. Two of the most crowded scenes, i.e. highest incidence of FG objects, are chosen to test with each sequence equating to roughly 300 frames. This equates to over 18000 segmentation label maps to evaluate.

3.4.1 Evaluation Metrics.

Three metrics are employed to assess the segmentation methods outlined in this chapter. The firsty of these is known as the Variation of Information (VOI) [120] which is an information theory based metric to describe the distance between two clusters. Consider two partitions X and Y of a set Z, split into disjoint subsets given as $X = \{X_1, X_2...X_k\}$ and $Y = \{Y_1, Y_2...Y_l\}$. If we then let $n = \sum\limits_i |X_i| = \sum\limits_j |Y_j| = |A|$, $p_i = |X_i|/n$, $q_j = |Y_j|/n$, $r_{ij} = |X_i \cap Y_j|/n$. Finally then, the VOI between two partitions can be mathematically defined as given in Equation 3.7.

$$VI(X;Y) = -\sum_{i,j} r_{ij} \left[\log(r_{ij}/p_i) + \log(r_{ij}/q_j)\right] \tag{3.7}$$

The Covering Rate (CR) [121] is an image segmentation metric to determine how well regions correctly overlap or *cover* target segments. It naturally extends the method of determining the overlap between two region R and $R^{'}$ given as $\mathcal{O}(R, R^{'}) = \frac{|R \cap R^{'}|}{|R \cup R^{'}|}$. The Covering Rate of a segmentation S by another segmentation $S^{'}$ can then be given as Equation 3.8. Both VOI and CR evaluation functions are applied to the standard label output segmentation maps.

$$C(S^{'} \to S) = \frac{1}{N} \sum_{R \in S} |R| \cdot \max_{R^{'} \in S^{'}} \mathcal{O}(R, R^{'}) \tag{3.8}$$

The Jaccard distance, conversely, is applied over the binarised layer output for BGFG and TBG extraction, where the Jaccard distance (JaccD) provides another measure of dissimilarity between two sets and evaluates the extracted BGFG region. This isolated background region is unrecoverable without FG context. To calculate the Jaccard Distance between two sets A and B, determining the Jaccard Index or *intersection over union* is the first step required. Thus, the Jaccard Index is given as $J(A, B) = \frac{A \cap B}{A \cup B}$ and JaccD can be simply computed as shown in Equation 3.4.1. Human ground truth label maps are utilised and the evaluation results are presented as multiple graphs in Section 3.4.2.

$$JaccD = 1 - J(A, B) \tag{3.9}$$

3.4.2 Results

An illustrative example of the challenge faced when recovering underlying background regions of FG objects, using segmentation methods *sans* context, is provided in Figure 3.5. This example utilises an indoor surveillance scene as it is a controlled environment showing people walking through a concourse. The colour video footage is recorded using static surveillance cameras, capturing data ideal for the development of an initial working solution. The image shows the region based EGB method failing to recover the background due to the presence of FG pedestrians, where it instead extracts them as individual regions. The SLIC segmentation method also, predictably, follows this path. An example ground truth label map of what we would like to recover is given too. We can now highlight our proposed background recovery algorithm performance on the same test image, which is provided in Figure

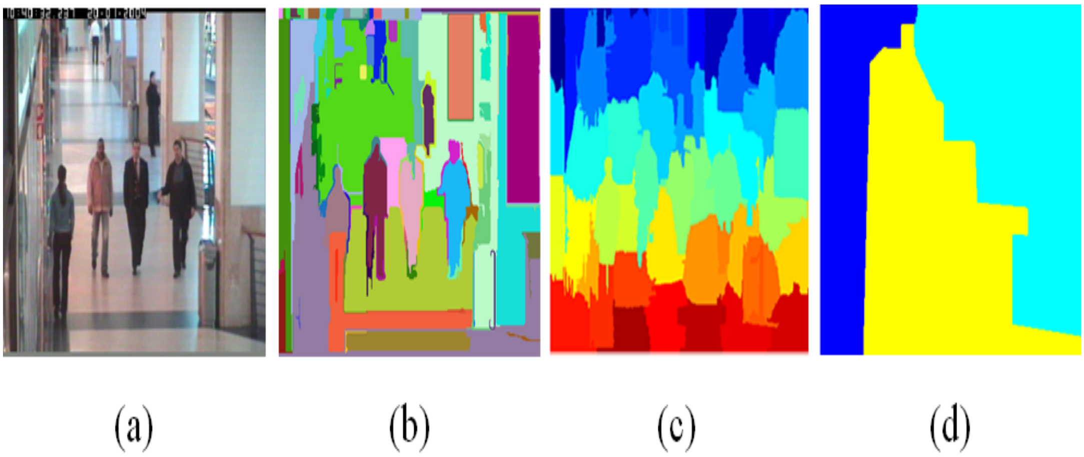


Figure 3.5: Image (a) shows a crowded scene from a static surveillance camera in a shopping centre as our input case [11]. Image (b) is an optimal EGB segmentation of the input image. Image (c) is a typical SLIC output label map, for an approximately desired size of 50 superpixels. Image (d) is a human ground truth example for the input shopping centre view.

3.6. The recovered background is clearly more contiguous in nature and closer to the desired label map when contextual knowledge is included. The most prominent BG region where FG objects are likely to appear is overlaid onto the original input image, further demonstrating the application of the method for background recovery. The binary layer can only be extracted due to the knowledge of FG objects and provides an enhanced scene understanding. The same type of BG extraction effect for a crowded urban scene is additionally shown in Figure 3.7, for experiments undertaken using the Oxford dataset. We can now quantitatively examine overall performance for the method proposed.

Figure 3.8 shows each segmentation methods VOI score over the selected datasets. Figure 3.9 and Figure 3.10 present the CR and JaccD scores respectively, for each segmentation method variation over the chosen crowded sequences. It is clear that the inclusion of FG context has a beneficial impact towards our goal of background extraction from cluttered scenes. In all cases presented this is shown by the proposed method scoring better than competing methods when contextual information is incorporated. Furthermore, the additional smoothing term has a positive gain in all occasion, as shown in Figure 3.10 for the Caviar dataset. Overall, these results can be summarised into a concise illustration, highlighting the performance gain offered by exploiting FG context. The percentage improvement for our best proposed method versus best competing segmentation algorithm score, for each dataset, is shown in Figure 3.11.

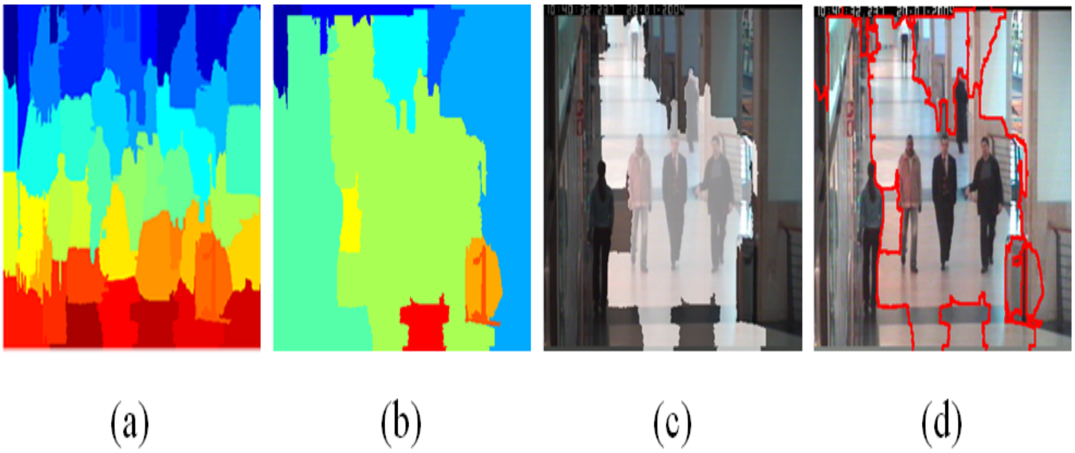


Figure 3.6: Image (a) shows the output SLIC label map from Figure 3.5, which serves as the input to our algorithm. Image (b) is the output of our proposed method incorporating contextual information. Image (c) is a binarised BGFG layer, overlaid onto the original RGB test image. Image (d) the original RGB input image overlaid with the contextual region segmentation boundaries, showing a large central BG region where FG objects tend to exist.

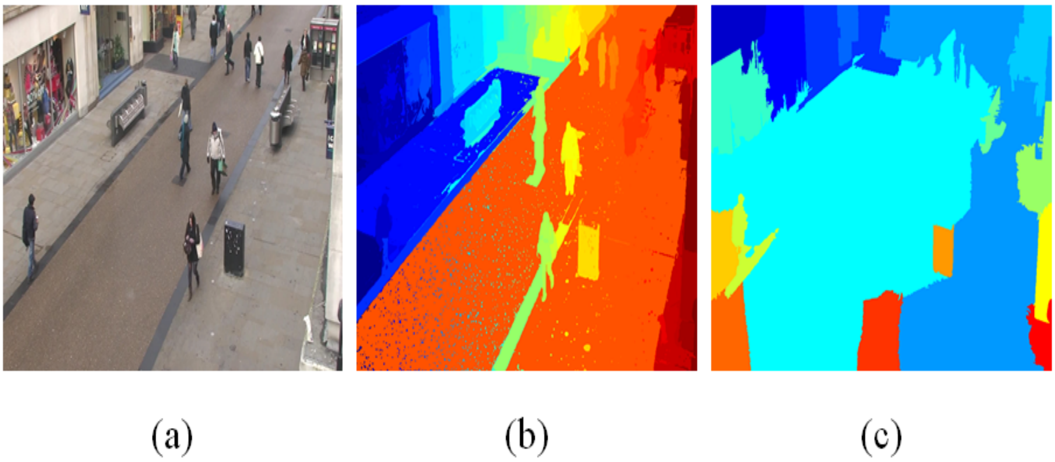


Figure 3.7: Image (a) is a static surveillance image from the Oxford dataset. Image (b) is a segmentation via the Efficient Graph Based algorithm, where pedestrians are clearly extracted as FG objects. Image (c) is the output from the contextual segmentation algorithm described in this chapter. It mostly extracts the underlying region while avoiding explicitly isolating people as separate regions.

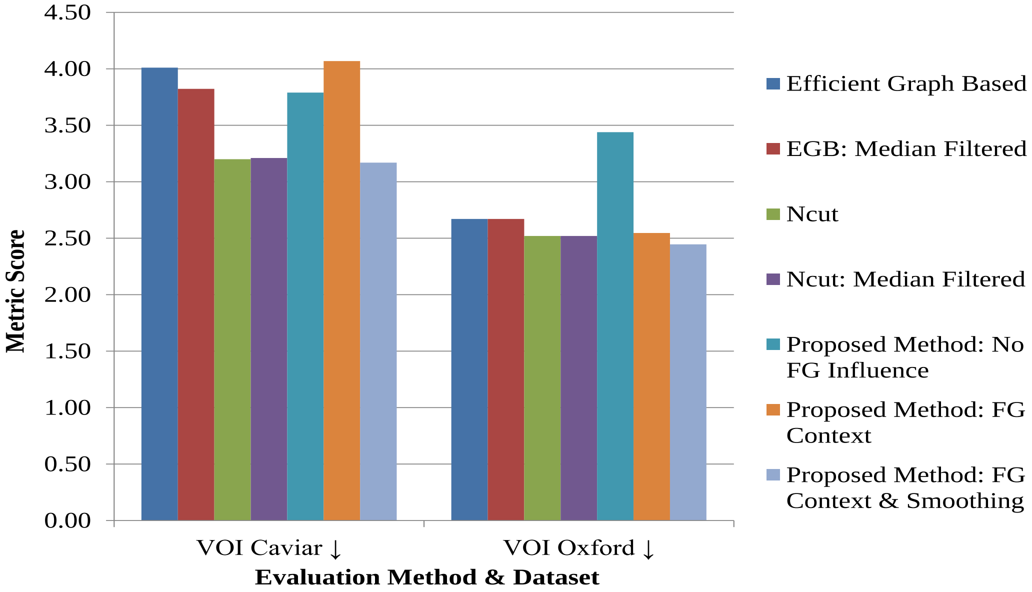


Figure 3.8: Evaluation on Caviar and Oxford dataset using VOI metric showing median averaged values. The arrows indicate where each metric should ideally be converging, e.g. downward arrow suggests minimising VOI is preferable.

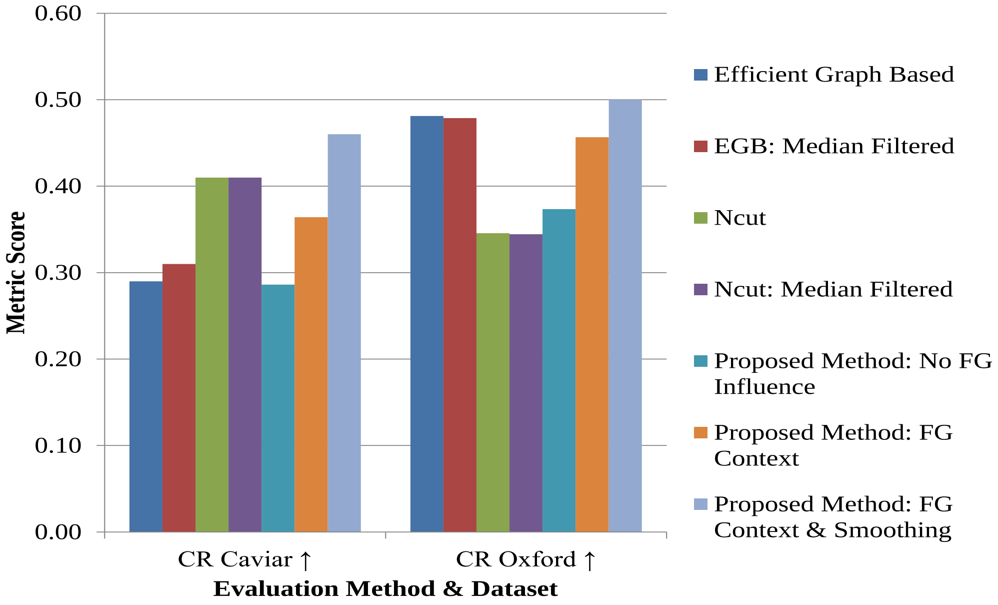


Figure 3.9: Evaluation on Caviar and Oxford dataset using CR metric showing median averaged values. The goal is to maximise the Covering Rate via segmentation.

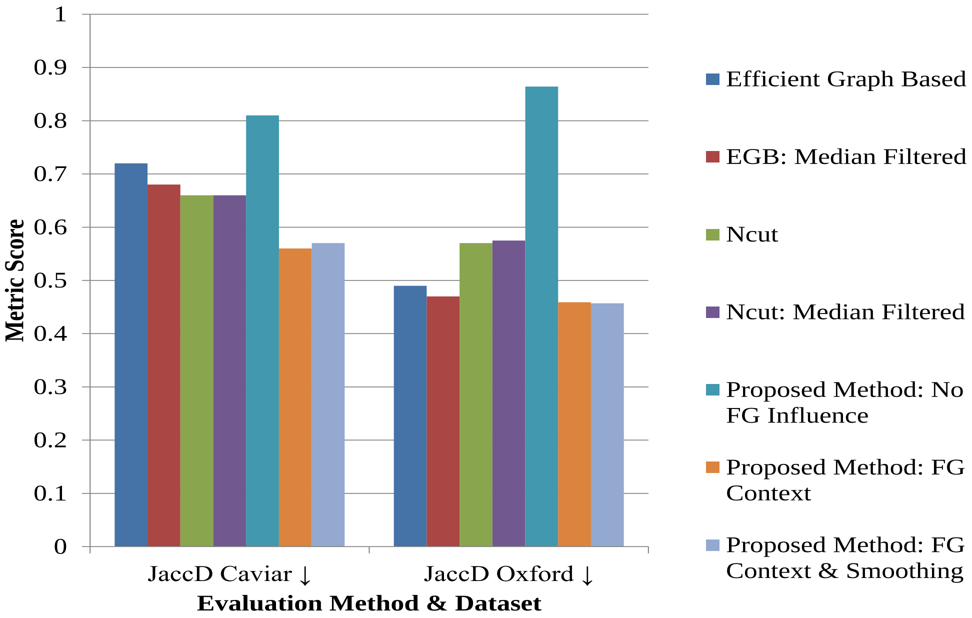


Figure 3.10: Evaluation on Caviar and Oxford dataset using JaccD metric showing median averaged values. The goal is to minimise this distance metric via the FG extracted binarised regions.

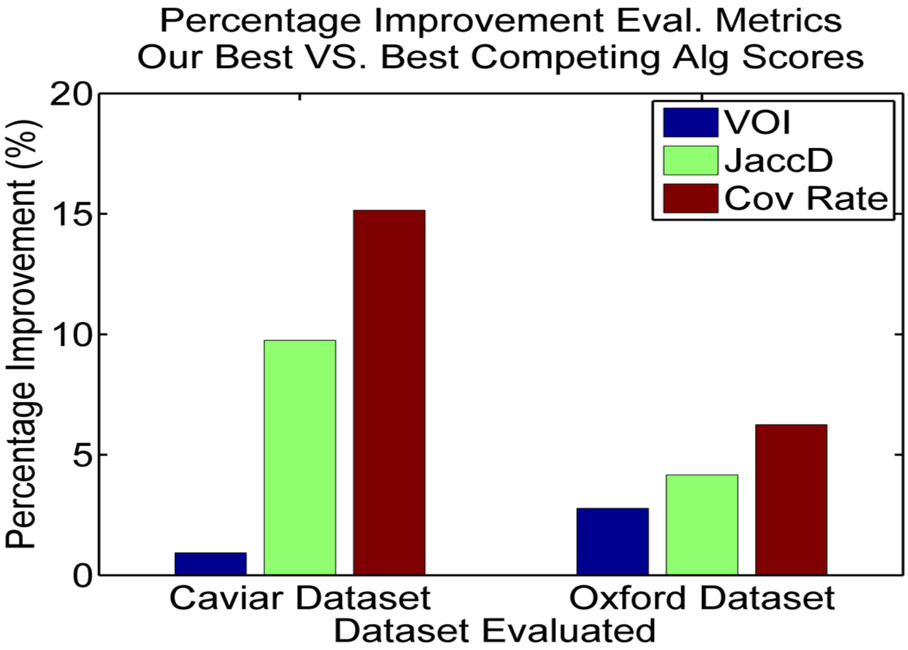


Figure 3.11: The bar chart presented shows the overall percentage improvement gain across the evaluation metrics employed for each dataset used. The values are generated by calculating the percentage change from the best competing segmentation method versus our proposed contextual segmentation algorithm.

3.5 Conclusion

This chapter demonstrates through comprehensive experiments that using available foreground context improves the segmentation process in crowded scenes, where the goal is to recover underlying background regions. The inclusion of FG context and smoothing improves the measures of VOI and CR cf. all competing methods. There is a negligible difference between using only FG context and incorporating smoothing for JaccD. The main benefit of our approach is that the algorithm is purely data-driven and only uses information available in the scene. Further, it delivers an underlying scene-structure that is unrecoverable without the inclusion of foreground context.

Given the overall EngD project aim is to enhance ATR performance with colour and thermal band sensors, there is a notable absence of infrared imagery in this chapter. The main reasoning behind this is the majority of related works on image segmentation and scene understanding are developed using colour imagery, which this chapter naturally extends. If the extracted scene structure or embedded foreground knowledge is present in the colour imagery, it will surely be present in corresponding modalities from the same static surveillance viewpoint. Thus, additional scene understanding gained from colour video can be applied to the 24 hour surveillance problem if used in conjunction with an additional thermal sensor.

This chapter lays the foundation for such an idea to be explored, which the following chapter addresses by incorporating key elements of the outlined segmentation process with an additional sensor modality. We aim to exploit the embedded foreground knowledge from colour band segmentation using our algorithm, followed by contextual knowledge transfer to obtain thermal target classification.

Chapter 4
Multimodal Object Classification via Contextual Foreground Regions

Chapter 3 discussed a scheme to recover background regions in cluttered urban scenes by exploiting the presence of foreground objects. The extracted segments are embedded with elements of foreground knowledge from this process. We explore ways to utilise the information gain for classification purposes in surveillance based applications. This brings us to the present chapter where we apply contextual segmentation principles to a multi-modal surveillance environment.

We present a novel method for classifying objects in a static-cam surveillance scenario using colour-thermal imagery. The solution is applicable to the 24*-hour surveillance problem and relies on exploiting scene-specific foreground context to determine regions of interest, leading to robust object detection. Moreover, a unified Bayesian framework is employed allowing the graceful exchange of contextual information across modes. This overcomes the dependence on individual sensor management as illumination conditions vary with time, a frequently occurring issue in outdoor surveillance when day transitions to night etc. In other words, we can use the knowledge transfer across modalities to classify thermal signals as specific types of target, without a trained infrared object classifier.*

An additional aim of employing the closely bounded segmentation and contextually guided system is to potentially circumvent the need to construct a trained infrared target classifier. The work contained within this chapter was presented at the 4*th IMA Mathematics in Defence Conference, 2015.*

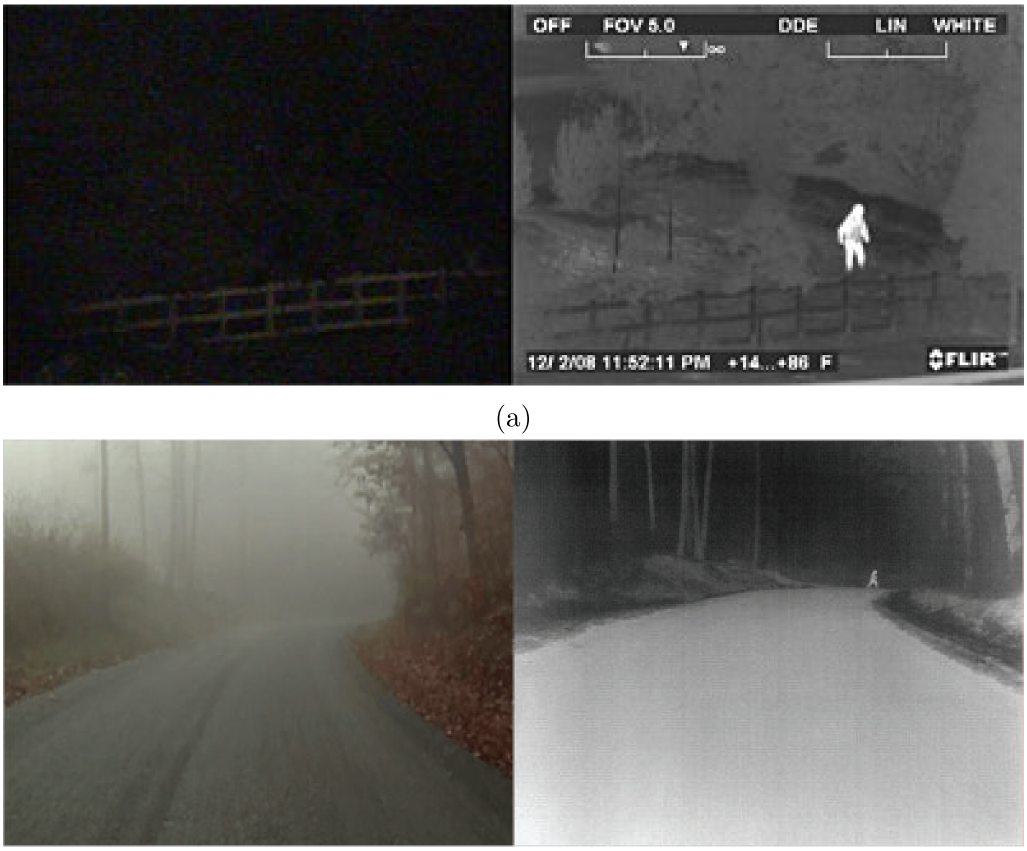
4.1 Introduction

Utilising video camera technology for surveillance and monitoring applications is commonplace and evident across a spectrum of sectors. This includes government agencies in a defence capacity to civilian usage, such as bolstering security around the home etc [122, 123]. Traditionally, surveillance systems would be reliant upon a human operator to perform effective object recognition and scene comprehension tasks. An operator would have to perform this task across a multi-camera system, requiring alertness in order to be efficacious. Whilst the capability of human vision is yet to be surpassed by computers, a potential weakness in the described system is the human-element.

Recent advances in computer vision have addressed this issue by creating detection algorithms that relieve some of the burden from operators [124]. The research field of computer vision produces algorithms that, ultimately, aim to replicate or improve upon the human-visual system. These algorithms should be capable of exhibiting human-like performance but at computer-like speeds. This is the main ambition computer vision, as a whole, hopes to realise and while a great deal of work has been directed towards creating such algorithms, it remains a long way off due to the inherently difficult nature of the problem [125].

The task of effective object detection and classification in computer vision is complex due to a variety of factors. Such challenges include the wide range of object appearances under differing poses, positions and scales exhibited in the real world. Another difficulty is dynamic illumination which affects how objects appear under varying light conditions [126]. Suppose traditional surveillance events are likely to occur during both day and night-time, in these cases it is imperative that corresponding detection algorithms can operate in these circumstances. Of course, it is dependent on what constitutes a *traditional surveillance event.* One such definition of a surveillance event is an observed action that is out of the ordinary, or a cause for concern, in the context of the scene [127].

For the purposes of this research work we consider the context of such an event to be an outdoor environment where objects within have no direct or limited control over illumination. Furthermore, image data would be collected from appropriate day and night sensors in a static-cam set-up. Possible surveillance scenarios include a covert reconnaissance mission or a border control environment [123]. A day sensor is essentially any modern colour camera. In contrast, a night sensor has to be capable



(a)

(b)

Figure 4.1: Two pairs of colour-thermal imagery illustrating the benefit of IR sensing. Image (a) presents a night-time surveillance scene where a colour camera cannot effectively perform. A thermal sensor, however, shows the presence of a person very clearly [13]. Image (b) illustrates an LWIR sensor piercing through a foggy scene clearly showing the presence of a hot body in the far-ground that cannot be seen in the corresponding colour imagery [1].

of *seeing* in the dark and be unaffected by varying light conditions. An example highlighting the benefit of thermal imagery in low light level conditions is presented in Figure 4.1.

As we have previously discussed, TIs are sensitive to EM Radiation in the IR domain and meet these requirements. However, IR sensors are under utilised compared to optical-band sensors. This is determined by several factors. Although there has been a notable reduction in cost for producing a quality TI, prices remain significant and present a barrier to wider use, where the price of a reasonable commercial TI system has dropped from roughly €50000 to €3000 today over the last 6 years [128]. Moreover, a price barrier coupled with night-time imaging capability places thermal sensing as a pursuit of mainly military and governmental organisations [13]. State-of-the-art TIs supply rich and textured imagery of target scenes, providing effective surveillance capability at night or through fog and smoke. The focus of this work will explore methods to address target acquisition and classification in 24-hour surveillance type imagery, using information gathered from a colour-band and LWIR camera.

4.1.1 Motivation

If we consider the knowledge that TIs are relatively uncommon there is a clear knock-on effect within computer vision research. Namely there is a corresponding lack of research available focusing on automatic target recognition methods in the thermal domain, *relative* to the colour-band domain. As we reviewed in Chapter 2, detection and classification methods rely on access to hand labelled training data for building a system with reasonable performance. Given the relative obscurity of the subject area, such curated sets of thermal data are not easy to obtain and usually involve significant upfront costs to generate. This leads to the specific problem of classifying objects in thermal imagery, without a trained classifier, which is addressed in this chapter. We hypothesize that inexpensive object classifications can be obtained in thermal imagery, via contextual knowledge transfer from colour domain detection information. Trained colour-based detectors are used to build FG context along with a recent segmentation method to create FG regions, corresponding to FG object activity. Foreground region knowledge acts as scene context and is exploited in the thermal domain.

Our goal is to show this approach is suited to a 24-hour surveillance problem within certain constraints, where we can employ and exploit the vast research available for object detection in colour imagery. This allows a TI to be employed effectively without the burden of training additional classifiers for IR images etc. We will demonstrate this using a publicly available dataset and a self-collected, colour-thermal dataset for the purposes of the experiment. The methods used to build FG context are explored in Section 4.2 and how this is transferred across modalities is discussed in Section 4.3. Following this are details of all experiments carried out in Section 4.4, including data acquisition and reported significant results in corresponding subsections. Lastly a summary of findings is presented in Section 4.5.

4.2 Utilising Foreground Context

Generally the methodology presented contains three central elements to complete the stated aims. The first task is to create FG regions for a scene given detection information, using both thermal and colour data. Following this is the extraction of features or *blobs* in IR imagery. Lastly a domain knowledge transfer for the determined FG regions allows the extracted thermal features to be classified, without a trained classifier. In essence this is a short summary of the entire process. This section will only focus on the underlying mechanics of building FG regions.

To approach the framed problem, foreground regions must first be constructed from observations. Two superpixel-based segmentation methods are considered to achieve this. The more complex of these, explored fully in Section 4.2.1, originates from a recent algorithm utilising FG context to segment improved background regions in scenes showing dense object clutter [23]. This algorithm is modified to incorporate additional thermal image information used in the experiment. The second method employed presents a simpler approach towards the creation of FG regions corresponding to areas of activity within a scene, which is described fully in Section 4.2.2. Let us explore each method respectively.

4.2.1 Modified Background Recovery

The algorithm presented by Rodger et al. was discussed fully in Chapter 3 [23]. It extracts the underlying area for FG objects in an image, where said objects are namely people. Given the recent exploration of this algorithm we shall only discuss how it is modified towards the current problem at hand.

To achieve our goal we adapt the background recovery algorithm in the following ways to suit our multi-modal surveillance problem. Firstly, the computation of a dissimilarity matrix using B_{dist} needs to incorporate pixel information from thermal imagery. This is easily achieved by simply including the additional IR channel information. We can determine a normalised histogram H using all pixels bound within each superpixel, for information channels, c. The normalised histogram for each channel is given by H_C . Considering we are dealing with colour-thermal data, c will be of the form $RGBT$. Obviously RGB relates to colour and is interchangeable regarding chosen colour space, whereby T is thermal infrared data. Thus, the distance between two superpixel distributions, SP_x and SP_y, can be given as Equation 4.1:

$$D_H(SP_x, SP_y, C_{RGBT}) = B_{dist}(H_C(SP_x), H_C(SP_y)) \tag{4.1}$$

where D_H is the dissimilarity (or distance) between each superpixel, calculated for each information channel and multiplied through. As previously mentioned, the resulting distances between superpixels is used to populate a dissimilarity matrix BD for each SP relative to every other SP, as described by the pairing $(x, y) \in A(SP_i)$. The creation of BD matrix forms the basis for region merging and is a key process, hence the need to incorporate an additional channel for IR.

We are interested in more than one object class while using this adapted segmentation scheme, namely cars and people. A capable detector should, therefore, be utilised to this end. We choose Aggregate Channel Features to detect pedestrians and SubCat to provide vehicle detections [129, 130]. These detectors are needed to create corresponding object trajectories, achieved by implementing a simple tracker that links detection points between frames. The subsequent tracks are then used to influence the region merging process giving mapped FG areas. Lastly, the final modification from the prior art involved focuses on the background-foreground layer computation. This binary array is adapted to provide probabilities over the FG region, so that pixel (x, y) is not solely binary but instead exists in the range $[0, 1] = \{x, y \in \mathbb{R} | 0 \leq x, y \leq 1\}$.

We deem the adaptation as an Aggregate Foreground-Background (AFGBG) map. Originally the binary map is determined every frame with only some information being carried forward to future segmentations. However, the adapted method accumulates each layer over the whole process so that each pixel location in the map is added throughout the sequence. If we let I be the binarised FGBG image array, then the complete output Z, from adding each binary layer for a sequence of n frames, is given by Equation 4.2:

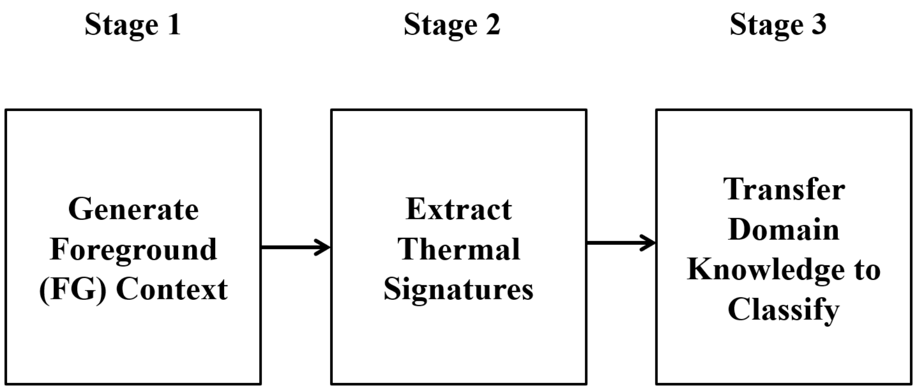

Stage 1
Stage 2
Stage 3
Generate Foreground (FG) Context
Extract Thermal Signatures
Transfer Domain Knowledge to Classify

Figure 4.2: This diagram outlines the major steps of the algorithm. At Stage 1 the foreground context is generated, via superpixel variance and trained detectors. Stage 2 identifies thermal targets in the scene using a feature extraction algorithm. Finally, Stage 3 employs a Bayesian framework to transfer context from colour to thermal domain, allowing thermal signatures to be classified.

$$Z = norm(\sum_{1}^{n} I_n) \tag{4.2}$$

where the usual rules of matrix addition apply. To elaborate, each element in I is added to the corresponding element in I_n, while output Z must share matrix dimensions with those added. Array Z is normalised so array elements exist over $[0, 1]$. This approach differs in that previously, only a small amount of information from subsequent frames was carried over for the final segmentation smoothing stage. Instead we retain the whole layer and effectively stack them over time, allowing a more complete picture of the AFGBG layer to be garnered from accumulating each array. Ultimately this region map corresponds to an objects likely area of activity, or conversely an objects inactivity, which will be our basis of classifying thermal signatures via contextual foreground regions.

4.2.2 Observing Superpixel Variance

By contrast the following algorithm description is relatively simpler from that explored in Section 4.2.1. It relies on observing the variance of image pixels, within superpixel patches, over a video sequence for both colour and thermal imagery. The key underlying assumption is that high variance corresponds with foreground object activity, for a stationary surveillance scene. The key processes in the method are presented in a flow diagram in Figure 4.2.

The most active superpixels are determined for each modality by segmenting a

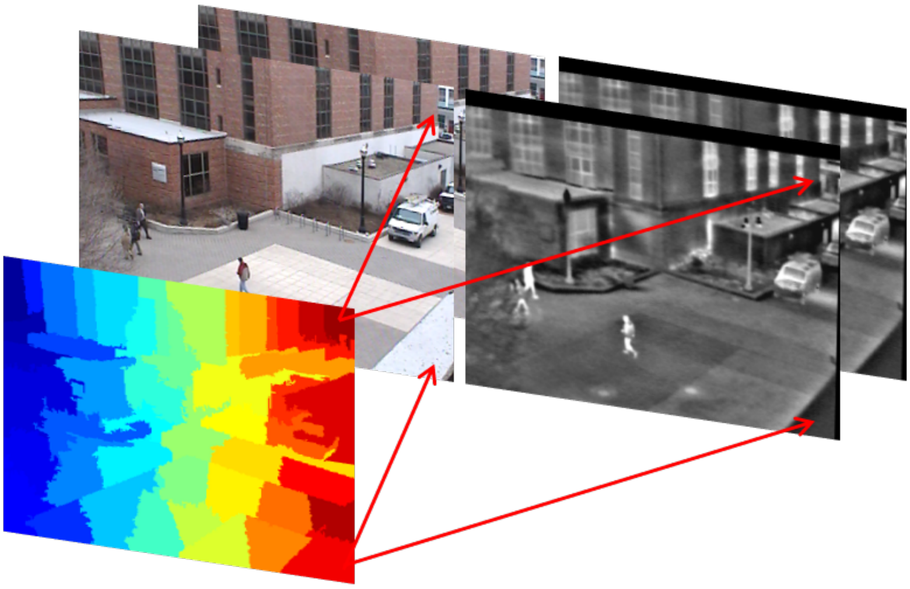


Figure 4.3: The framework takes initial Superpixel Representation in colour domain and observes the underlying pixel content in both modalities. For each superpixel the pixel variance in IR and colour images is then calculated for every frame.

frame of the colour image source, then effectively *tracking* the variance of pixels underlying each SP in *both* modalities. This concept is shown graphically in Figure 4.3 where a scene is observed through a superpixel representation.

Once the most active superpixels, in terms of pixel variance, are obtained these can be merged to form a SP variance layer indicating the presence of FG objects. The SP variance layer obtained from the scene shown in Figure 4.3 is presented in Figure 4.4.

Referring to Figure 4.2, the remaining step to complete Stage 1 is to accumulate detections over the sequence and create confidence maps for each object class detected. The detection bounding boxes are mapped to the SP variance layer and if it lies with a region of activity (high variance), it is convolved with a Gaussian kernel to give a spread of confidence scores in range $[0, 1]$ for each bounding box region. If the detection lies in a low variance region, the kernel confidence score is halved for the mapped bounding box. This process leads to a AFGBG region map after normalisation. Again, this AFGBG layer has areas corresponding to object class activity, akin to the AFGBG layer given in Section 4.2.1.

To describe the method mathematically we must again consider the initial superpixel set. The first colour image in a sequence of length t is superpixellated using the SLIC algorithm, returning set n_i regions $SP_i = \{SP_{i,1}, ..., SP_{i,n_i}\}$. Given

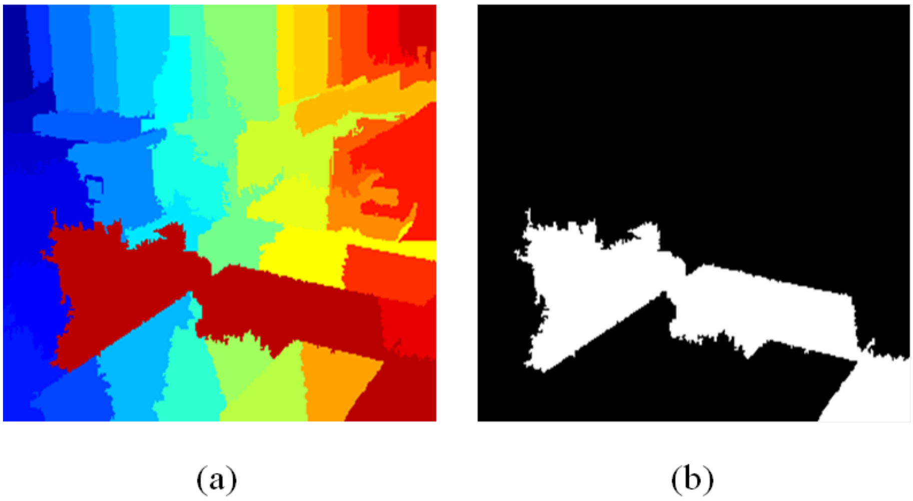


Figure 4.4: Merge highest varying superpixels across both modalities to form a variance layer. Image (a) shows the most active superpixels being merged. Image (b) is the binarised version of this to obtain the variance layer. The key underlying underlying assumption is high variance corresponds to FG activity.

we only need one superpixel representation to proceed, $i = 1$ in this case. The resulting label map can be used to observe pixel variance for *both* colour and thermal imagery, given both modes of imagery should share dimensions. Pixels $(x, y)_t$ underlying each superpixel will then have an associated label, allowing pixels belonging to each superpixel to be read and stored in vector form. Let this be shown as $(x, y)_t \in SP_n = A_n^t$, where each vector of pixels A is determined by the set label n and image sequence length t. For example, if the number of superpixels desired was 50 for an image sequence of length 100, then we would have 50 vectors of image pixels for each of the 100 images per modality or channel. Thus we can calculate the variance occurring for each superpixel per image by Equation 4.3:

$$G_c^t = var(A_n^t) \tag{4.3}$$

where G_c^t is the array of n superpixel variances through a sequence of images of length t and the variance function var is defined as Equation 4.4:

$$\frac{1}{N-1}\sum_{i=1}^{N}|B_i - \mu|^2 \tag{4.4}$$

where B is a variable vector of N scalar observations and μ is the mean of B.

From this it is easy to determine the superpixels exhibiting the most variance across the whole scene, by simply employing Equation 4.4 again but this time to array G for each channel. This produces a vector of length n for each channel c, indicating variance observed within each superpixel over the entire sequence.

Let the vector of SP variances per channel be deemed SPV_c. The next step is to perform a simple mean threshold to eliminate any SPs that are showing low to no signs of change over time. The final process combines information from the vector SPV_c to form a predicate for merging superpixels. If we let X_c be a set of integers corresponding to the most varying superpixels for each channel c, then we can obtain a set of integers M that indicate which superpixels should be merged to form the FGBG region map. For instance, if only 2 channels are used for this method, equivalent to IR and the blue colour channel for example, then the set of superpixels to merge can be obtained by performing a *union* of sets, presented as Equation 4.5:

$$M = X_1 \cup X_2, \qquad X_{12} \in n_1 \tag{4.5}$$

where n_1 is set of integers for number of superpixels used. By merging superpixels in set M we create a SP variance layer from simply observing pixel value variance through time, this layer is shown in Figure 4.4. However, no object class information is incorporated with the SP variance map unlike the method explained in Section 4.2.1. To achieve this a trained object classifier in the colour domain must again be utilised.

The basic principle is to store and accumulate any object class detections to incorporate with the FGBG layer. For each detection made per colour image, a corresponding bounding box will exist. These detection bounding boxes are then mapped to the SP variance layer, illustrated in Figure 4.5 where active regions are shown as white and low variance regions are black. The bounding box area can be convolved with a Gaussian kernel to provide it with a confidence score. For each convolved detection box, if it lies within an active region the scores remain unchanged. However, if the detection lies out-with the active regions then the scores are simply halved. This process remains the same regardless of class and is to reflect the uncertainty of making a detection in a region that has not exhibited much variance. The aggregation of these mapped and convolved detections, for each object class, leads to the creation of a FGBG map after a defined length of observation. Again, the score aggregated map must be normalised so array elements exist in the range $[0, 1]$.

Both methods explored generate a confidence region map for object classes, based upon observations over time. The problem is posed as a static surveillance scenario where classifications in the thermal domain can be achieved without having a thermal classifier. This is achieved by utilising a trained detector in the colour domain

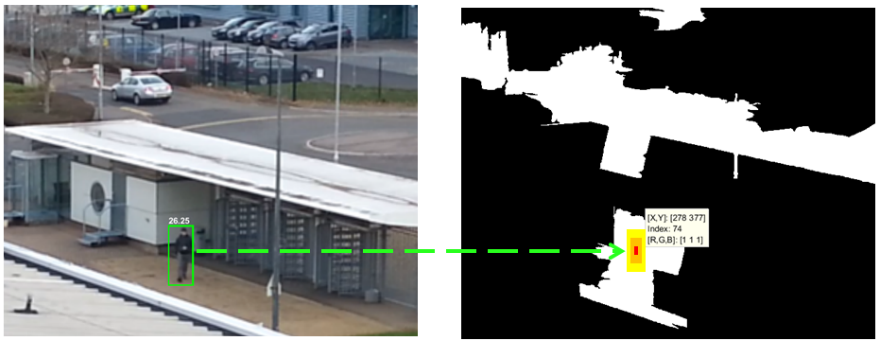


Figure 4.5: Person detection bounding box is mapped to SP variance layer and convolved with Gaussian kernel to aggregate and build confidence map, which is the Foreground-Background layer. The resulting FGBG layer will be utilised to classify thermal signatures without a classifier at a later stage.

and transferring the observed knowledge to be fully exploited. The methods outlined above provide the mechanism to effectively build up *prior* information that can be carried forward to help detect thermal blobs, the details of which are discussed more fully in Section 4.3.

4.3 Transferring Knowledge Between Sensor Modalities

The classification process for thermal features is a two stage process, operating under the assumption that the previously available colour signal is now unavailable. Thus, all objects have to be identified in the thermal domain without a trained classifier. The first step is the extraction of thermal features from target imagery, which are then classified via a Bayesian framework using the obtained FGBG maps. The algorithm of choice to achieve thermal feature extraction is Maximally Stable Extremal Regions (MSER) [131], specifically the *VLFeat* implementation.

4.3.1 Thermal Feature Extraction

Referring to Figure 4.2, acquiring thermal signatures is Stage 2 of this algorithm. The core idea of using MSER to obtain features in LWIR imagery towards the task of pedestrian classification is not a new one, as evidenced by previous works [132]. However, we present a completely different approach to prior art in a similar context. The MSER algorithm is ultimately a feature / blob detector that aims to extract stable connected components for level sets of a given image. The *stability* of regions is determined by how much variation exists within each binarised component. The algorithm is often chosen due to its simplicity, robustness, while it works for low resolution imagery and has small computational cost.

One drawback is that it can often be hard to fine-tune. The MSER algorithm is utilised to extract numerous thermal blobs for each thermal image frame in a sequence. Each blob is classified as belonging to one object class, or not, by utilising previously obtained AFGBG maps for each object class in a Bayesian scheme.

4.3.2 Bayesian Framework

Referring to Figure 4.2, developing a contextual classification scheme is Stage 3 of the process. Bayes theorem is employed to compute the probability of a detected object to exist given the condition of its surrounding region probability [133]. Using the commonly given Bayes proportionality $P(A|B) \propto P(B|A)P(A)$, we can express our posterior probability as $P(Obj|R)$. The object is the extracted MSER blobs and the region R is determined by the earlier obtained AFGBG map, relating to each object class. This is shown in Equation 4.6.

$$P(Obj|R) = \frac{P(R|Obj)P(Obj)}{P(R)} \tag{4.6}$$

where $P(Obj)$ is the *prior* information for objects, $P(R|Obj)$ is the likelihood which usually expresses a prediction model for given data and the denominator $P(R)$ is simply a normalising constant.

The prior information is the output AFGBG layers from the methods outlined in Section 4.2, as it is an aggregation of detection observations shown as a confidence map. In other words it is the state of knowledge before the experiment to classify thermal features. Thus, to calculate the posterior $P(Obj|R)$ for an MSER blob belonging to an object class, or vice versa, it is a simple case of choosing a likelihood probability for each object class and summing probability values from the relevant prior AFGBG maps.

Every classification task is treated as a binary problem where an MSER blob can either be an object such as a person, or not a person and two probability values must then be computed to reflect this. The correct FGBG map created from specific object class detections must be used to determine the corresponding posterior accurately, where the FGBG prior is then inverted to calculate the probability of a thermal blob *not* being an object. Given it is always a binary problem a flat or uninformative likelihood of $P(R|Obj) = [0.5, 0.5]$ can be chosen, at least initially, to produce results that are free from bias. An uninformative likelihood essentially tells us that a thermal blob belonging to an object class or vice versa is equally likely.

Lastly, the probability of prior $P(Obj)$ for each MSER blob is simply an average of corresponding FGBG pixel values at locations where extracted thermal features appear. Once probabilities for every MSER have been obtained and normalised, the highest value indicates if the blob is a specific object class or not. Posterior values for every frame are stored along with corresponding MSER bounding boxes for future evaluation using human ground truth data.

4.4 Evaluation

The described methods for classifying thermal signals in a surveillance scenario, without a trained classifier, are tested using two datasets. For each dataset the FGBG confidence maps are constructed using three differing lengths of observation, for every defined object class. The output probabilities for each object class and associated blob bounding boxes are then used to calculate classification accuracy via human ground truth information.

4.4.1 Data Acquisition

For the experiments undertaken two datasets are employed. Firstly, a publicly available colour-thermal dataset is obtained from the OSU Color-Thermal Database - the OTCBVS dataset [38]. This set of LWIR and colour imagery is already registered spatially and temporally, whilst also being a static-cam surveillance scenario containing a sparse number of people as foreground objects. This provides an excellent platform to initially test our proposed methods. Example imagery from the OTCBVS set is illustrated in Figure 4.6. The second dataset was collected using a low-cost colour camera embedded in a mobile device and a state-of-the-art TI produced by Thales, the Catherine MP [20]. This set contains lots of clutter and an additional foreground object of interest, the car. It represents more of a real-world urban surveillance scenario in contrast with the relatively *sanitised* OTCBVS dataset.

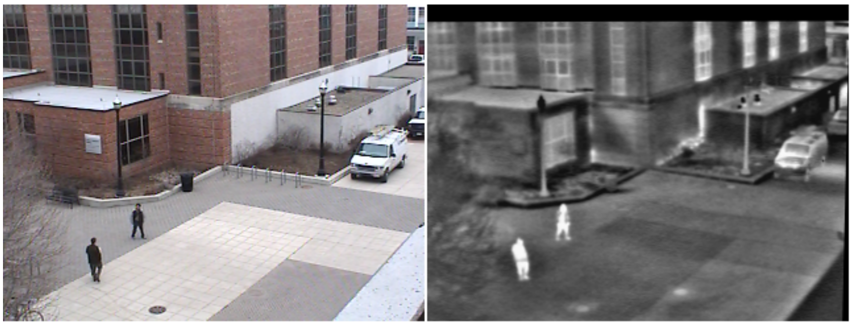


Figure 4.6: Registered colour-thermal image pair from OTCVBS dataset showing two people in a sparsely populated urban environment [14].

The Catherine MP LWIR uses an integrated detector cooler assembly which comprises a 640 × 512, 20μm pitch QWIP array, sensitive to long wave infrared radiation at wavelengths of 8μm to 12μm at a frame rate of 100-Hz. Given the accompanying colour camera has a frame rate of approximately 30-Hz with a larger spatial resolution, any image data collected using these two sensors has to be processed before being useful for experiments. To elaborate the colour and thermal imagery must be registered spatially so they show the same image plane, as well as temporally registered where objects within both images exist at the same point in time.

To enforce spatial coherence between Catherine MP imagery and colour imagery, a straightforward image alignment technique is employed. Manually selected control points identifying common features in both images are chosen, which allows a transform matrix to be computed via a geometric mapping process. This matrix is then utilised to perform the spatial transformation of imagery. It is acknowledged that much more sophisticated and automatic options exist to do this between multimodal image sets, but for the purposes of this experiment they are not deemed absolutely necessary given a reasonable level of accuracy can be obtained to effectively map FGBG layers from one modality to the other. Temporal registration is a much simpler issue that only requires the difference in frame rates as a ratio. For the Catherine MP to colour camera situation this ratio is ≈ 3.334, so for every 3 frames traversed in the colour sequence, 10 frames have elapsed in the LWIR sequence. This is enforced to create a real world colour-thermal dataset with high-quality, LWIR thermal imagery. An example of this image set is provided in Figure 4.7.

4.4.2 Test Conditions

For both datasets the image sequences are split into two sets. The first set is used to build FGBG / prior maps for each object class, using *3* different lengths of ob-

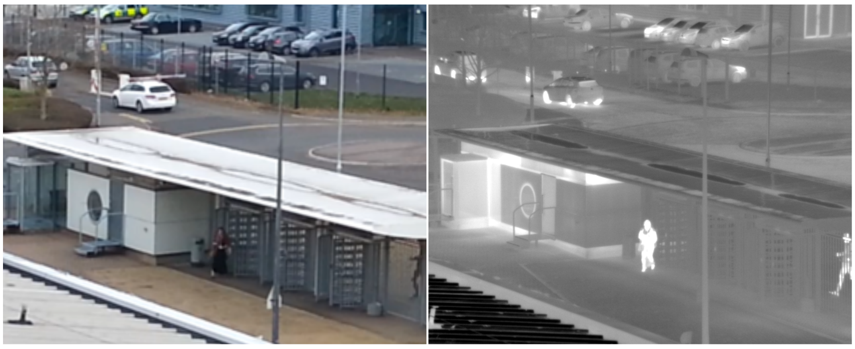


Figure 4.7: Registered colour-thermal image pair from self-collected dataset showing people, cars and clutter in a populated urban environment.

servation in terms of number of images. The second set of images is used to test the algorithmic framework for classification, under the assumption that the colour information is useless (i.e. night-time), where only thermal imagery is utilised. The test sets are manually ground truthed for both datasets. The OTCBVS contains only people as foreground objects while the collected dataset contains both people and cars. Bounding boxes are generated for these object classes.

The algorithm parameters, such as MSER variations, are kept constant throughout. The only change is the number of superpixels double from 50 SPs for the OTCBVS data to 100 SPs, accounting for the larger resolution present in the collected dataset. A flat likelihood of $[0.5, 0.5]$ is used for all experiments, where probabilities and FGBG maps are generated for each object class via both methods. The length of observation for the OTCBVS set is presented as multiples of 1000 images, with a test set of roughly 1000 images. The rationale behind the size of the test set is purely down to the number of images in the chosen sequence from the OTCBVS dataset. The intention is to keep a test sequence completely isolated from any observations used to generate the prior knowledge. The higher quality imagery, in terms of thermal information, self-collected dataset uses multiples of 250 images and a test set of approximately 350 images for presentation. In both experiments, data points are generated at 0.5 increments of the observation set size. For example, the first point on Figure 4.8 occurs at 0.5 multiplied by 1000 images, meaning the observation set was 500 images in size.

4.4.3 Classification Performance

The bounding boxes generated from MSER classification are used to evaluate the overall system accuracy by calculating overlap with ground-truth object bounding boxes. The metric used is *Accuracy* which is calculated using Equation 4.7:

$$ACCURACY = \frac{\Sigma TP + \Sigma TN}{\Sigma TP + \Sigma TN + \Sigma FP + \Sigma FN} \tag{4.7}$$

where TP is a true positive, TN is a true negative, FP is a false positive and FN is a false negative. For this experiment, a TP is where an extracted MSER has a greater probability of being an object (car/person) and also has a greater than 50% overlap with the correct object class bounding box. A TN is where the dominant probability deems the thermal blob as *not an object* and there is less than 50% overlap with ground-truth bounding box. A FP is obtained when the highest MSER probability classes the blob *as an object* but there is not sufficient overlap with ground-truth boxes. Lastly, a FN is when an ample overlap exists between MSER and ground truth boxes, but the dominant probability deems the blob as *not an object.* This overlap convention is defined in Equation 4.8 and is a common evaluation approach for detection problems [63].

$$a_0 = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \tag{4.8}$$

In Equation 4.8 $B_p \cap B_{gt}$ is the intersection of MSER bounding boxes B_p and object ground truth bounding box B_{gt}, while $B_p \cup B_{gt}$ is their union. All variations of accuracy results, for each algorithm, over the OTCBVS and self-collected dataset are presented in Figures 4.8 to 4.11 .

The obtained accuracy results show clearly that the simpler method of superpixel variance, combined with ordinary detection information in the colour domain, is an effective method to classify thermal features. Or to a greater extent, it is a more robust and accurate method for building the AFGBG prior maps used in the MSER classification process, when compared to the complex *Modified Background Recovery* approach. This is true regardless of the object classes examined. It also appears that in most cases the length of observation does not appear to have an overwhelmingly positive effect. Initially this seems counter-intuitive as increasing FG object evidence should provide a more accurate platform to classify thermal signatures.

However, there may be several non-obvious factors that mean this is not the case. For instance, the adapted background recovery will tend to propagate errors through a sequence if the scene is not densely populated. Thus the longer observation length will only compound this problem. Furthermore, longer sequences to compute superpixel variance through a scene can lead to key areas falling out of

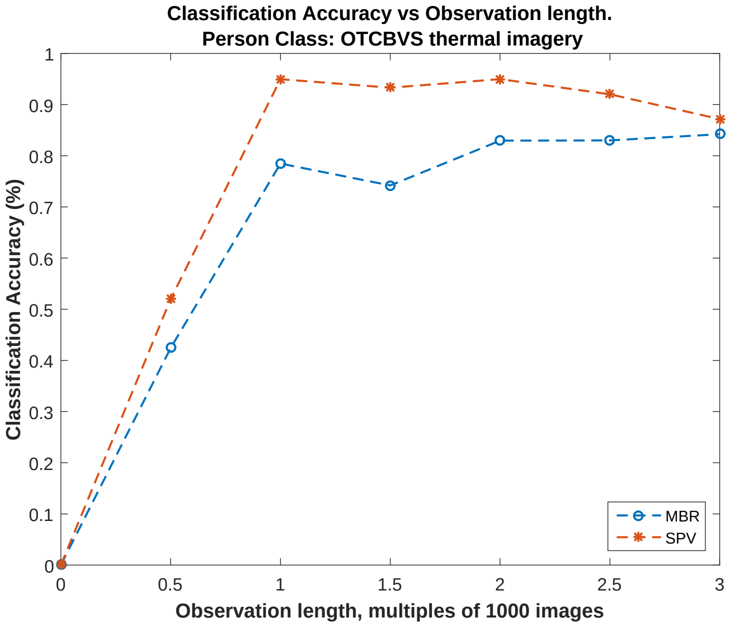


Figure 4.8: This graph shows accuracy results for person classification over OTCBVS dataset for two competing methods. Prior maps were generated over varying image sequence durations, as indicated on the $X - Axis$.

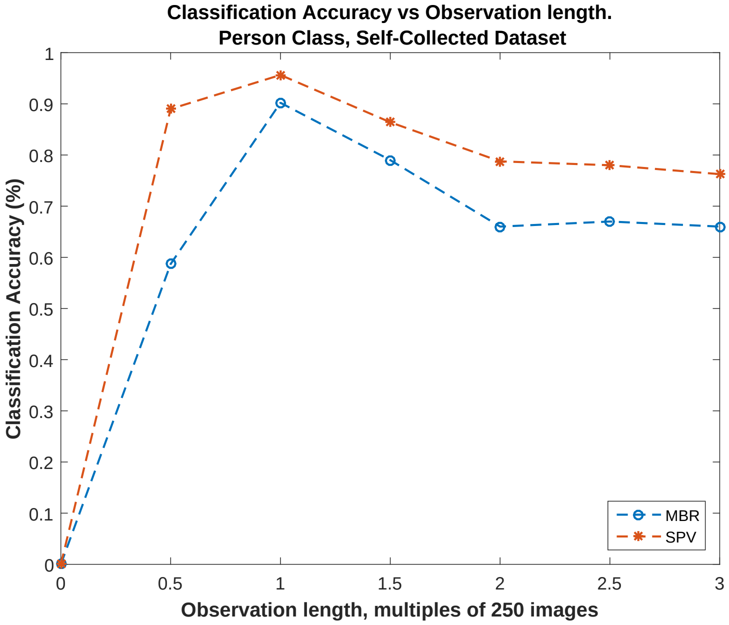


Figure 4.9: This graph shows accuracy results for person classification over a self-collected multi-modal image dataset.

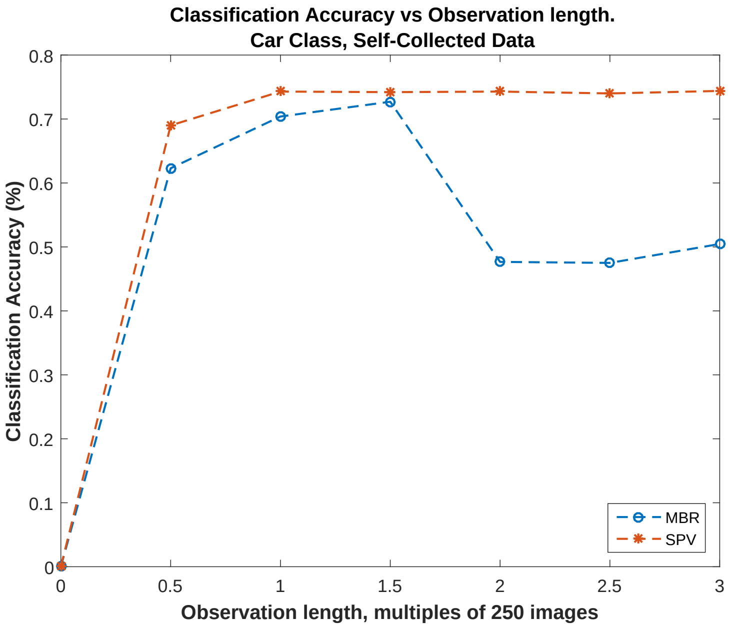


Figure 4.10: This graph presents the accuracy results for car classification over a self-collected multi-modal image dataset.

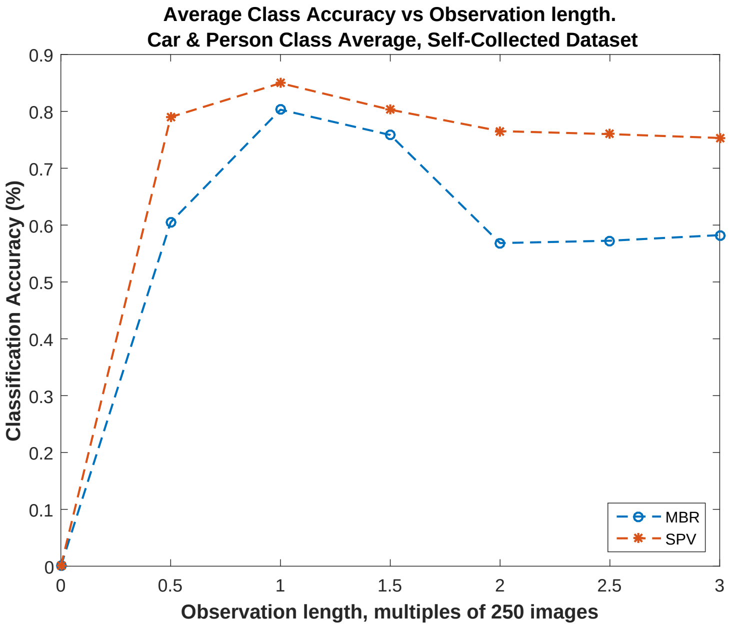


Figure 4.11: This graph gives an averaged class accuracy for both methods described, with results generated over the self-collected dataset.

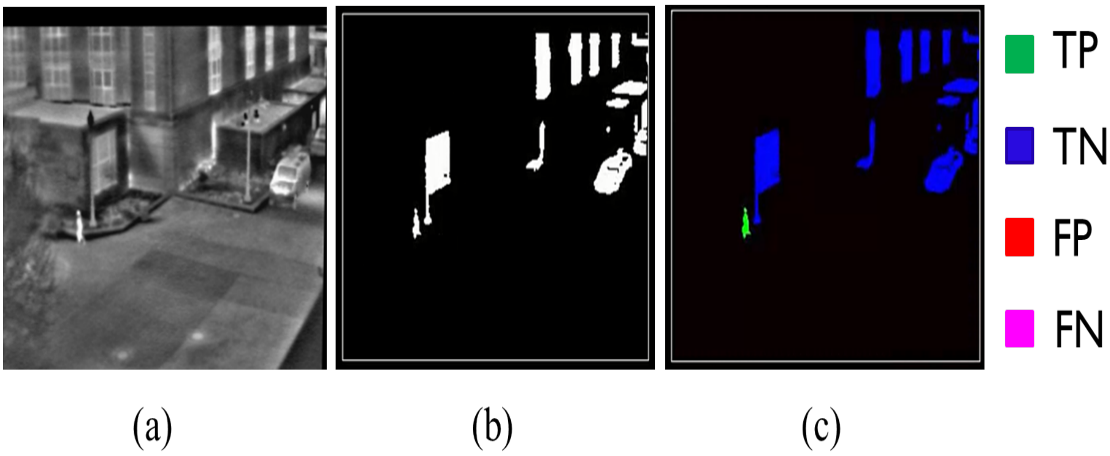


Figure 4.12: Image (a) is a typical surveillance scene captured in LWIR. Image (b) is the thermal signatures our work will classify to provide object information. Image (c) illustrates the output classification from our contextual classification scheme, with the key on the right hand side. The person is clearly identified as green pixels, with background regions shown in blue.

the SP variance map later on as the scene changes. In any case it is clear that while improvements can be made the SP Variance method provides a more accurate AFGBG map, feeding directly into the Bayesian classification scheme for extracted thermal features. An illustrative example of thermal feature classification using the proposed superpixel variance algorithm is presented in Figure 4.12.

4.5 Conclusion

The main hypothesis put forward in this chapter examines the transfer of contextual knowledge across modes enables object classifications, without using a trained classifier. Ultimately the experiments carried out have shown this to be true. Two methods for generating foreground contextual knowledge are presented and the output of these serve as input to a Bayesian classification scheme, which mediates the knowledge transfer between modes. Initially, a complex background segmentation algorithm is adapted to suit the problem posed, which in turn leads to a simpler solution utilising superpixel variance to determine regions of activity.

From experiments carried out over two colour-thermal datasets it is readily apparent that the simpler of these methods, combined with the Bayesian classification framework, is better suited to generate contextual foreground regions to aid multimodal detection. Despite identifying a potential solution for classifying objects in

LWIR imagery, the approach is quite limited in many ways. For instance, either method discussed requires prior observation periods from a static surveillance viewpoint using thermal and colour band sensors. Furthermore, the methods themselves are fairly convoluted and lack a certain simplicity that is favourable in deployment of real world applications. Lastly, the methods gain class specific information for targets using trained classifiers for colour imagery, meaning they are likely only as good as the classifiers overall accuracy. Thus, it seems very unlikely the algorithms outlined in this chapter will mature into a tractable solution.

At the time this work was conducted we partially responded to the growing advance of machine learning methods by circumventing the large effort required to create a labelled object dataset in LWIR. Such a dataset would allow the application of effective machine learning techniques to build a bespoke LWIR target classifier. If this could be achieved it would propose a much simpler solution for thermal object recognition which is a critical aim of the project. The successes of such machine learning methods has been well documented but mostly for colour imagery, so we can now explore the application of CNNs towards LWIR imagery and build a state of the art thermal object classification scheme.

Chapter 5
CNNs for LWIR Object Classification

In this chapter we present an end to end process to create an object classifier using CNNs for LWIR thermal imagery. After the reasonably complex solution proposed for thermal object recognition discussed in the previous chapter, we arrive at the conclusion that building a CNN recognition model will offer more advantages despite the significant upfront effort involved.

The subject matter and underlying theory is laid out before describing the thermal dataset creation in full. We report the collection of LWIR image sequences using the Thales Catherine MP and how this is transformed into a labelled, multi-object dataset. Preprocessing and data augmentation steps are also discussed at this stage in order to create an image corpus suitable for training a robust CNN classifier. Then we move onto CNN architecture design considerations and methods for hyper-parameter optimisation. Finally, we can assess and evaluate the CNN recognition performance using our labelled dataset, observing accuracy across our defined object classes. Furthermore, we also utilise recent network visualisation techniques to ensure CNN behaviour is as expected. Lastly we re-implement the trained CNN in a different deep learning framework and deploy a realtime solution, feeding in LWIR video sequences recorded using the Catherine MP for target recognition purposes.

To the best of our knowledge this research presents one of the first successful applications of CNNs to LWIR imagery for object recognition across several classes. Elements of this chapter were presented at SPIE Security & Defence, Electro-Optical and Infrared Systems; Technology and Applications, 2016 *[16]*

5.1 Introduction

Accurate and robust recognition capabilities have long been sought after in computer vision, where the ultimate goal is to develop algorithms approaching human level performance [134, 135]. Recent machine learning methods have solved challenging image problems across a number of applications, such as object classification and image segmentation. Convolutional Neural Networks are one such method offering a potential route to achieve intelligent processing systems, comparable to human vision for tasks such as object recognition [136]. In this chapter we utilise a deep learning framework with CNNs to create an object classifier for longwave infrared thermal imagery.

Infrared sensing technology is widely adopted in the security and defence domain due to its night vision capability [137–141]. This is possible due to the nature of thermal radiation, where the LWIR portion of the Electromagnetic Spectrum is dominated by heat emissions. Passive sensors designed to be sensitive to this waveband, from $8 - 12\mu$m, can thus offer persistent surveillance capabilities [20]. Enhanced situational awareness for end users can also be gained if the data stream is intelligently processed in some fashion, offering decision assistance. Intelligent processing methods offer the additional benefit of limitless *attention span* compared to human operators, who would typically suffer performance degradation at repetitive or stressful tasks [19, 142, 143]. Modern machine learning methods for computer vision applications are one route to achieving this, which is explored and advanced in this chapter.

Deep convolutional networks have demonstrated an excellent ability to learn representations of data, allowing the generation of highly descriptive features. Crucially, these features are generated automatically via the machine learning process, moving away from traditional human crafted features that often under perform by comparison. In principal, improved and increasingly generalised features can be obtained through careful network design and training schemes to avoid overfitting [144].

The convolutional network has shot to fame in recent years despite existing for decades [145]. There are several factors contributing to this rise in prominence. Firstly, computing infrastructure and hardware has greatly improved over approximately the last twenty years. This has enabled the collection and distribution of sufficiently large image training sets necessary for scaling deep learning models. Furthermore, the proliferation of advanced accelerated processing hardware allows training schemes to operate on much shorter timescales. However, the largest influence illuminating the power of CNNs was the well-known results for the 2012 ImageNet challenge [146]. In the words of Andrej Karpethy, this competition is effectively the *"world cup of computer vision"* [147]. Thus, when a CNN entry dominated other algorithms by a large margin, in terms of accuracy, the field quickly took notice. Convolutional networks have since become widely adopted in academia and industry across many applications [48].

While this revolutionising approach is extensively utilised for colour band imagery, i.e. visible light, there is surprisingly few examples of CNN applications for thermal band data by comparison. Several aspects may contribute to this situation. For instance, high quality thermal imagers are mostly found in the security and military domain due to the high price barrier, meaning the necessary imagery for training a CNN is restricted from the outset. Nevertheless, we take full advantage of the advanced machine learning capability offered by deep networks to train our own architecture. The tuned CNN is designed to classify objects of interest in LWIR imagery.

We collect and ground truth a sufficiently large LWIR object database to enable successful training of the CNN. The imagery is effectively preprocessed and balanced by using suitable data augmentation techniques. Experimental results indicate excellent performance across all object classes after training. Further examination of the internal network structure also demonstrates the training scheme is appropriate and we explore the feature space to illustrate network behaviour accordingly.

Ultimately, we show that deep networks can be employed in a very similar fashion for high level computer vision tasks using thermal band data, as already shown for colour band data. This realisation could lead to future developments of intelligent automated systems in security and defence, which typically rely on efficient processing of thermal imagery.

5.2 Related Work - Analysing Thermal Data

The application of machine learning using CNNs for Automatic Target Recognition (ATR) tasks is an active and well documented topic of research. Despite the high performance offered by deep networks and the importance of ATR in defence scenarios, the field is relatively unpopulated by CNNs for thermal based data. We shall examine methods for ATR processes that utilise convnets, as well as alternative techniques, specifically in the infrared domain. For the purposes of exploring classification strategies, it is assumed the target acquisition process is complete and we are dealing with a set of potential candidate *blobs* extracted from a thermal image. The image blob is usually pre-processed via a simple set of operations, which we will describe for our method in Section 5.3. Following this procedure, relevant and descriptive features can be extracted allowing machine learning based classification methodologies to be examined.

5.2.1 Constructing A Classifier

Generally, object classification strategies prior to the widespread adoption of deep convnets rely on a two stage process. Assuming we have sufficient candidate targets available, the first task is concerned with feature crafting. The aim is to extract relevant attributes from the data that best captures discriminatory aspects of the signal, whilst reducing redundant information. Ideally, these features should be generalisable to the task at hand and offer suitable performance when only a subset of descriptors are available [43]. For instance, in an ATR scenario we may only have access to half an image of a land vehicle due to clutter obscuration, meaning an incomplete set of features will be generated for classification purposes. In cases such as these it is hoped a subset of features will still offer sufficient levels of performance. Lastly, is it usual practice to organise extracted features into a vector or $2-$dimensional feature image which ultimately is a representation of candidate targets [44].

After successfully assembling a feature construct comes the second stage in the process, requiring a classification scheme. If we let the described feature vector be known as X, composed of n feature instances, then $X = \{x_1, ..., x_n, \}$. Using this terminology allows us to express a candidate target, i.e. a transformed image blob, as a feature vector X, where the goal in ATR is to determine what object class, C, a potential target belongs to. Furthermore, in the context of machine learning based classifiers an additional choice must be made with regards to the learning paradigm. For ATR tasks the problem tends to be well bounded and constrained so it is usual to see a supervised learning approach, where labelled training data is required to infer a predictive function. Thus, a classifier during training would require input pairs of data, formed of the feature vector X and a corresponding label Y. After training is complete a classifier will generate a prediction of Y for any new input feature vector X. The alternative to supervised learning is unsupervised learning,

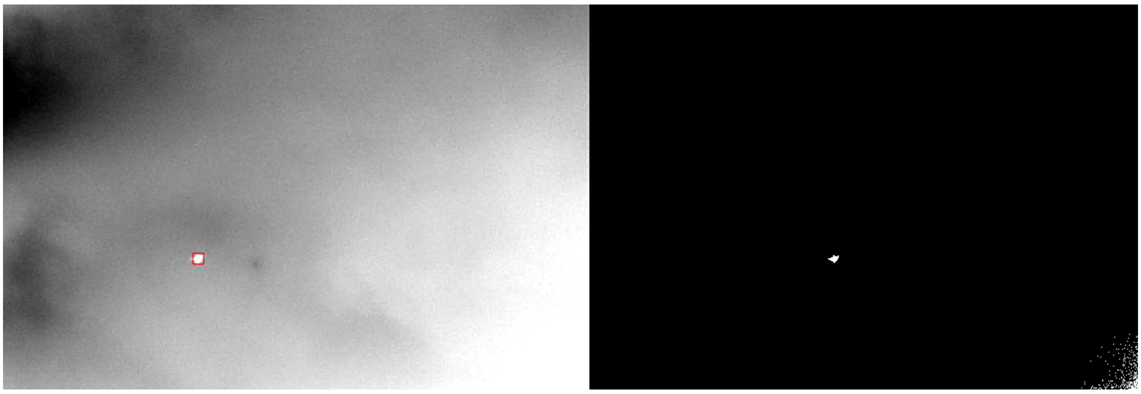


Figure 5.1: An LWIR thermal image with a UAV against a sky background is shown alongside its binarised counterpart, extracted using the MSER algorithm. The UAV is highlighted by a red target box.

which does not offer predictions of Y for input X but instead clusters the features accordingly. Unsupervised approaches shall not be considered here. The task of assigning an object class C to candidates based on a feature vector X summarises a target classification procedure [45].

5.2.2 Alternative ATR Schemes

Due to the 24-hour sensing capability offered by thermal imagers they enable valuable detection and classification processes in the security and defence domain. There are many possible options for performing object recognition tasks using thermal imagery, without invoking a CNN approach. For instance, a wide area search and surveillance system presented by Breckon et al. employs cascaded Haar classifiers to generate potential target candidates from an input thermal image [148–150]. The search windows returned via the cascaded Haar classifier are then confirmed or rejected as a target type using a secondary trained classifier. Note that the initial target generation uses Haar basis functions and derivatives to obtain features and the classifier is trained using AdaBoost. The secondary target confirmation classifier generates a feature vector using Laplacian filter responses over input candidate patches, which are then used to train an SVM classifier [151]. Both of these examples adhere to the two stage process for classifier construction outlined in Section 5.2.1.

Another popular method for feature generation in LWIR imagery is to first obtain *blobs* using a technique known as MSER [131] and extract discriminatory attributes from them. An example of a UAV in thermal imagery being highlighted by MSER is shown in Figure 5.1 to illustrate the usefulness of this feature extractor.

These features are then used to train an SVM classifier or similar algorithm. For instance, a low resolution pedestrian detector and classifier is shown in [152], while a person identification system for assisting *special weapons and tactics* teams is illustrated in [153]. Both example methods follow this general approach of hand-crafting features for classifier training. In these cases features are generated by applying various methods to MSER blobs to obtain descriptors, such as Discrete Cosine Transformation [154], Histogram of Oriented Gradients [74] and Integral Channel Features [155]. The object descriptors serve as input to a modified Random Naïve Bayes [156] or SVM classifier for training and deployment.

Obscured target recognition systems also utilise LWIR sensors and classification schemes to identify potentially hazardous buried targets. The works of [157,158] for instance employ a forward-looking longwave infrared sensor and create object descriptors via ocal binary patterns [159], among others. Again, hand crafted features are used to train an SVM classifier for predictive purposes towards explosive hazard detection.

It should now be clear that machine learning based recognition methods adhere to the outlined two stage process, consisting of feature crafting and providing input pairs $\{X, Y\}$ to a classifier of choice for training. The end result is effectively an optimised inference function capable of assigning a target class for new input feature vectors X. The small selection of example methods report reasonable performance for their respective tasks, some even present incredible performance of $\approx 99\%$ area under the curve [152]. However, despite the successful appearance of such methods there exists a significant underlying drawback.

Ultimately this general approach of hand-crafting discriminative features for specific tasks suffers from the human element involved, as we are always making a decision at some level of what we *think* makes a good feature for object classes. This is a time consuming process that actually limits the generalisable and scalable nature of ATR methods, which stems from the fact that humans cannot really know what makes a good feature at the *machine level.* By designing descriptors and passing them to a chosen classifier it is explicitly based on our knowledge and partial assumptions of how the human visual system performs ATR, but this may not always translate to corresponding levels of computer performance. Under the supervised learning paradigm we can now explore options where the manual feature generation process is removed, allowing the machine to craft them automatically.

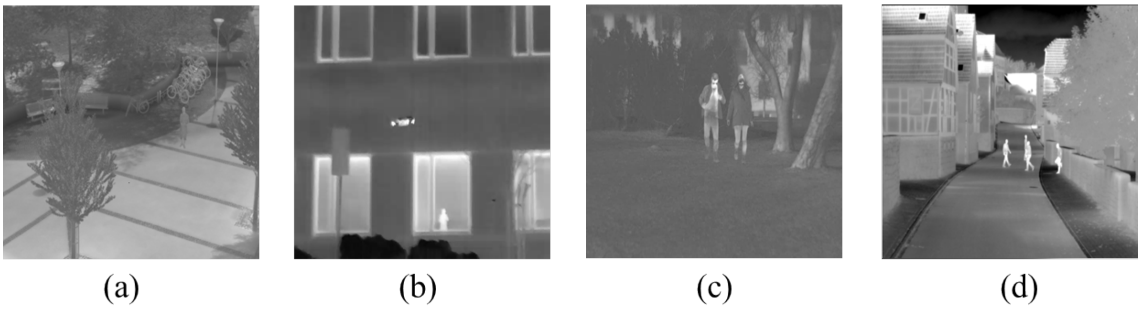


Figure 5.2: Different object examples in LWIR are illustrated in the public, labelled dataset provided by Berg et al. While it is likely a useful resource to develop algorithms the dataset contains only a handful of object classes relevant for defence applications, the human and quadcopter. Humans can be observed in images (a) - (d), while image (b) shows a quadcopter. Even so it is encouraging to see such a dataset in the public domain. Image source [15].

5.2.3 CNN Based ATR

We have touched upon several prior methods for ATR in the thermal domain that utilise manually generated features and typically an SVM classifier. Such methods are numerous as the approach proved reasonably successful over time. Given the prominent rise of CNNs and increased visibility over recent years, it would be easy to assume there is now a significant body of work employing CNNs for ATR tasks in infrared imagery. However, publications and research in this area is surprisingly sparse despite the intelligent recognition capability offered by CNNs. The underlying reason for this may be attributed to the lack of available, large scale labelled datasets for infrared imagery. A corresponding and well known dataset in the colour band by comparison would be Imagenet [146], containing over 1000 object classes over approximately 1.2 million images and is easily accessible to those with an internet connection. Ground truthed datasets of this magnitude are very rare in the infrared domain. One of the only comparable public datasets, produced by Berg et al. [15], contains humans and quadcopters which is useful for security applications, but it also contains horses and dogs which is less desirable from a defense standpoint. Example imagery from this dataset is presented in Figure 5.2.

Lastly, the lack of prior art may be down to the notion that CNNs must be trained over large scale datasets such as Imagenet. This assertion is refuted by the training of a CNN from *scratch* using a modest image corpus later on. For now, let us explore the handful of available CNN based ATR systems.

One of the earliest prior works employing CNNs for ATR with infrared imagery conducts an empirical evaluation of several methods towards this task, where they observe neural network based approaches generally perform better and are easier to implement due to the lack of manual feature engineering required [160]. They also note that while CNNs perform better than competing methods, they are routinely confused by examples that are simple for humans. This is to be expected given the small amount of training data available and lack of advanced techniques implemented in modern convnets, which would have limited how well their CNN generalised to the task.

The detection and recognition of buried targets in FL-LWIR has also been tackled by CNNs as shown in [161]. Curiously, although the authors try a multitude of CNN variations they discover the convnets cannot match performance levels of their baseline algorithm, composed of hand-crafted features, for the ATR task. It is left unanswered as to why this is the case but it appears to be a combination of network architecture and what CNNs actually learn. The variety of features used in the baseline algorithm appear to be very closely linked to physical characteristics of buried targets, whereas CNNs are ultimately relying on image appearance to extract features. This a rare case where a CNN fails to outperform a manually engineered classifier.

Another application of convnets can be seen in recognition of low resolution targets from near-infrared aerial imagery, which has proven feasible using a CNN model approach [162]. This example highlights a novel network architecture designed to classify challenging low resolution targets and it outperforms pre-trained networks as well as hand designed feature methods. Moreover, the classic computer vision task of pedestrian recognition is solvable using CNNs for ATR [16, 163, 164]. As assisted driver systems become more advanced they will require robust object detection and recognition systems able to work in varying illumination conditions, so it is unsurprising to see a CNN approach utilising far-infrared sensors [163]. The CNN developed for driver assistance is a lightweight CNN, as real-time computation is necessary in deployment, where the goal is to not only recognise pedestrians but to also identify potentially unsafe behaviour. They report encouraging results using a novel architecture consisting of a lightweight CNN connected to a boosted random forest classifier. Furthermore, the low-resolution pedestrian recognition system proposed in [164] is a natural extension of prior art by [152], where features are generated via CNN training instead of manually. The end-to-end model operates in real-time for LWIR video using a trained CNN for ATR, demonstrating excellent performance.

Lastly, the LWIR object detection and classification system presented in [16], which is also explored in Chapter 6, appears to be the most extensive CNN based work for LWIR imagery. To the best of our knowledge the CNN as demonstrated is capable of robust recognition for more object classes at various ranges and poses than other prior art in this field. We fully explore the details of the machine learning process, from constructing a balanced training set to feature visualisation, in order to train a robust, scalable CNN for ATR applications using LWIR input data.

All of the methods outlined in this section follow the supervised learning process, requiring labelled datasets for training purposes. Crucially though, the methods also benefit from automatic feature discovery during training. Removing the need for manual feature extraction not only saves time, but allows a network to best determine how classes should be differentiated. As we shall see, this is achieved by loss function optimisation over many iterations during the training phase.

5.3 CNN Creation

Supervised learning with convolutional neural networks for image recognition tasks is analogous to learning a new language, where the primary method of inference is examination of labelled images for different object types. The human brain will form connections to explain the new information by learning how to associate the visual data to the label. The new-found knowledge is reinforced and improved via testing to highlight areas of weakness. The end result is an accurate and generalisable knowledge within the brain, applicable to future recognition tasks.

Recent machine learning methods have captured the essence of this problem and provided a solution. In the case of object recognition for imagery, the state-of-the-art solution has been to employ convolutional neural networks with access to *enough* training data and sophisticated optimisation techniques. Accelerated hardware also helps but is purely optional to speed up the learning process. A neural network is essentially the recursive application of weighted functions succeeded by non-linear functions. A CNN is a special case designed to exploit image structure and a deep, feed-forward CNN is simply a stack of CNN layers incorporating downsampling and activation functions. The stack of layers gives rise to the term *deep* and the convolutional operations transform image inputs into feature response maps, or activation maps, whilst preserving the spatial structure of the image.

Once a convolutional network has propagated an input image through the CNN blocks and sufficiently downsampled the resulting feature maps, it will eventually reach a fully-connected layer (i.e. a traditional neural network). This final layer takes the output from the previous layer and is connected to all its neurons. Whilst the spatial information is finally lost at this stage the linear combination of weights provides very powerful and abstract responses, ultimately inferred from the training data. It is these generalisable inferences that we aim to extract from our thermal images. However, unlike colour imagery which is ubiquitous and accessible, the thermal domain is much more restrictive. Given deep neural networks require an appreciable amount of training data, the lack of labelled and accessible imagery is the first hurdle to overcome.

5.3.1 Dataset Generation

The first step to generating a successful CNN model requires access to a training set that is sufficiently representative of the problem domain. As mentioned previously this is particularly challenging to obtain for imagery collected using thermal imagers. It is even more difficult to gain access to human labelled thermal imagery suitable for supervised machine learning using CNNs. That is not to say no publicly available and ground-truthed thermal datasets exist, for instance the OTCBVS repository offers labelled LWIR scenes, enabling the undertaking of closely related research [14, 164].

While such provisions are sparsely offered for thermal data they share a common theme, the thermal data is of low quality. The major contributing factor for this is attributed to price. High quality TIs are very expensive as they are designed to capture as much of a designated portion of the infrared spectrum as possible, to a high degree of sensitivity, which is a non-trivial task to achieve. This challenge becomes even more formidable when we consider the LWIR band, where the optical efficiency of capture systems in this wavelength tend to be very poor. In other words, to obtain high quality thermal imagery we must use a high quality TI, which are scarcely accessible to the public or academic community.

Rich thermal imagery captured by such a TI would allow a robust CNN model to be trained for the more challenging problem of extensible object recognition, at various distances and poses. This problem is more akin to real world scenarios and such a recognition capability would be valuable. Thus, in the absence of such a labelled dataset we set out to obtain one. It requires a sufficient volume of video data to be collected using a high performance TI. Following this, important object classes for land defence and the image preprocessing stage is defined.

5.3.2 LWIR Dataset Generation

To recap, the Catherine MP LWIR is a state-of-the-art TI produced by Thales and is employed to collect data towards creating a suitable training set for our chosen target classes, which are explained further on in this chapter [20].

Acquisition: The first step to overcome is simply one of data gathering. The state of the art Catherine MP is employed to collect sufficient data in order to build a robust CNN. An integrated detector cooler assembly is housed within the Catherine MP and is comprised of a 640 × 512, 20μm pitch Quantum Well Infrared Photodetector (QWIP) array. The photodetector is sensitive to long wave infrared radiation at wavelengths of 8μm to 12μm at a frame rate of 100Hz. This TI is presented in Figure 5.3 deployed in a sensor platform. Crucially, the longwave thermal imagery acquired using this system will be of sufficiently high quality to enable the application of deep convolutional networks with thermal band data, for more challenging and real-world type images. This outlines the TI collection system.

Multiple terabytes of video footage were recorded using this TI, capturing a range of object classes from multiple poses, ranges and in different weather conditions. Targets were imaged at a range of a few metres to a few kilometres. The crucial aspect to address when building a dataset for machine learning is that the data must accurately reflect the real-world problem. This gives the chosen machine learning paradigm the best possible chance of determining an accurate *worldview* of the problem domain. Having gathered a large amount of video footage, a harness was written in MATLAB to traverse through each video sequence and crop out suitable target candidates from the frame. These image crops were stored as raw pixel values and labelled as the appropriate target class. The objects of interest for the experiment are defined as follows.

Desired Target Class: Surveillance and land defence is the driving force behind the objects deemed to be of interest, where we identify objects frequently observed in rural and urban environments. Thus, the target class is composed of five real objects and one null case, which we designate as the false alarm class. The real target classes are people, land-vehicle, helicopter, aeroplane and Unmanned Aerial Vehicle (UAV) . An illustrative example of each real target class is presented in Figure 5.4, demonstrating the LWIR thermal imagery contained within the training

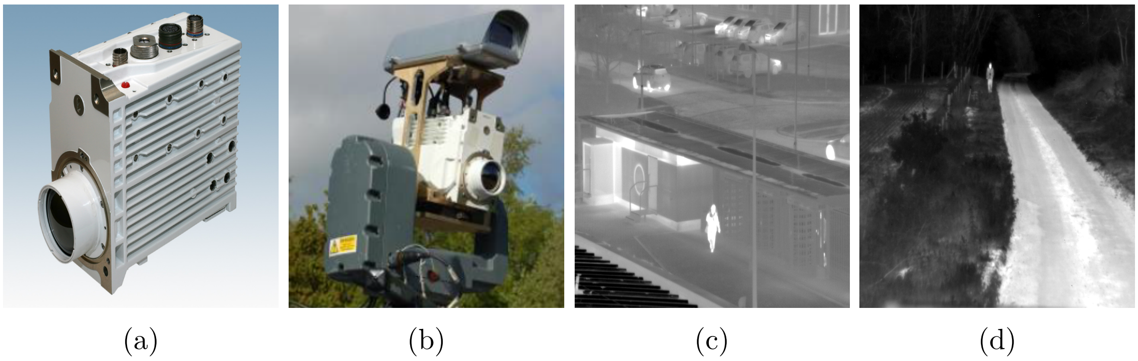


Figure 5.3: The Catherine MP LWIR variant presented as a single unit (a) and deployed on a multi-modal sensor platform (b). Two example LWIR images are provided in (c) and (d) to illustrate the appearance of the modality as well as highlight the quality of the sensor itself.

and test sets. The land vehicle class includes an assortment of many different types of ground based vehicles, such as a personal car, vans and construction vehicles. The false alarm class is comprised from an assortment of background clutter present in everyday scenes. This includes clutter items such as edges of buildings, clouds, foliage and patches of ground etc. Null cases such as these are typically problematic for object detection algorithms, especially in thermal imagery, where they tend to generate false alarms.

Examples of images labelled as null or false alarm are presented in Figure 5.5 to highlight the difference between real targets and false alarms. The most important aspect behind the inclusion of such a class, despite the possible conflict and confusion it may introduce to the CNN during the learning process, is the potential rejection power it enables. Target recognition algorithms akin to trained CNNs are ultimately designed for deployment in an overall vision system and our case is no different. We briefly explore a real-time deployment of such a system in a later section.

The ability to reject false alarms by building such cases into a network is a valid strategy and one that has seldom been explored. An alternative to this approach would be to recover real probabilistic values from the network, unlike the pseudo-probabilities offered by softmax output, to generate confidence scores from predictions [165]. Low confidence scores could then be thresholded in a similar manner. Now we have collected suitable data and defined the objects of interest we need to process it for successful training and validation.

Preprocessing & Data Augmentation: Following the designation of target classes the next stage requires the creation of a balanced training and test set, com-

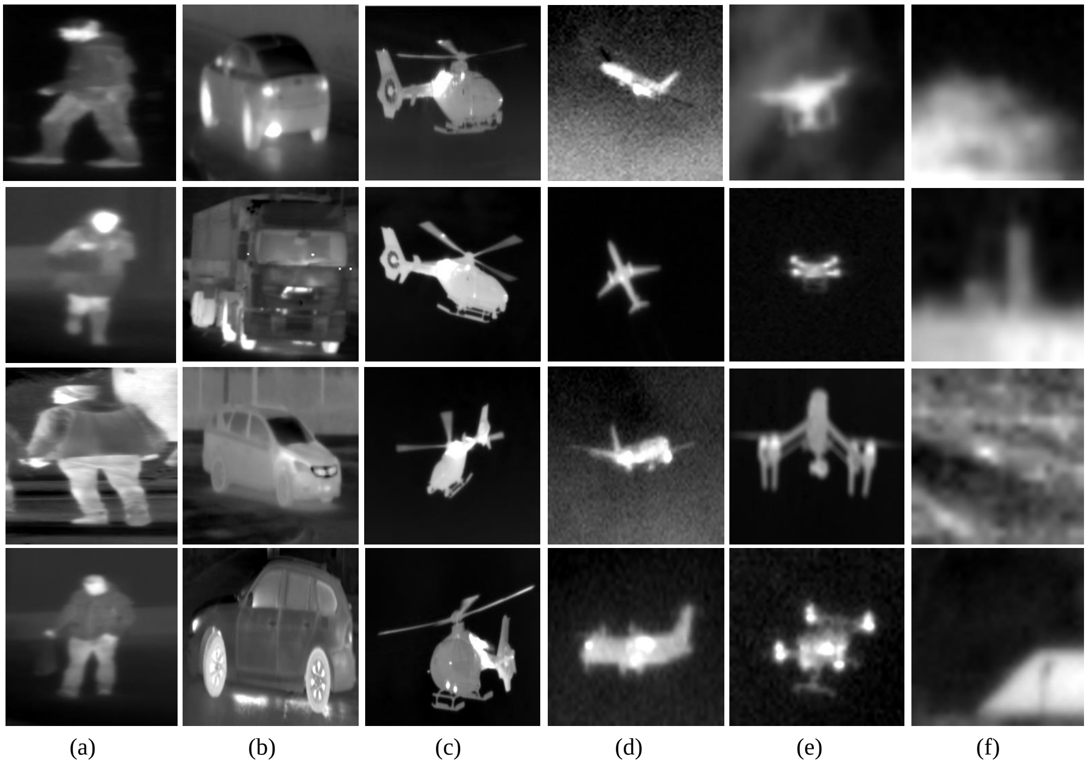


Figure 5.4: Training examples for each object class, cropped from Catherine MP LWIR imagery. Instances shown in column (a) highlight people from various poses. Land vehicles are observed in column (b), showing not only different pose/viewpoints but also intra-class variation. Instances of helicopters can be found in column (c), with aeroplanes present in (d). Lastly, column (e) illustrates UAV examples with column (f) highlights various false alarm instances. Image source [16]

posed of the outlined objects present in LWIR images. The Catherine MP outputs 14bit video data and relevant target classes can be extracted from the sequences as an individual image crop. For each object crop then, the image dimensions will vary per example. This is due to not only the dynamic nature of object appearance and pose, but also the distance said objects appear at. The box sizes of each crop vary accordingly. The employed CNN model requires input images to be of a fixed size and 32*pt* floating precision.

We elect the image dimensions to be fixed as 256 × 256 based on the following reasoning. Influential prior works that created CNNs for object recognition also resized images to 256 × 256, as the dimension allows a large number of downsampling operations to be performed if required. Given that building a successful CNN for thermal imagery is an unknown at this stage, we posit that such a large image dimension allows more opportunity to downsample or perform other preprocessing

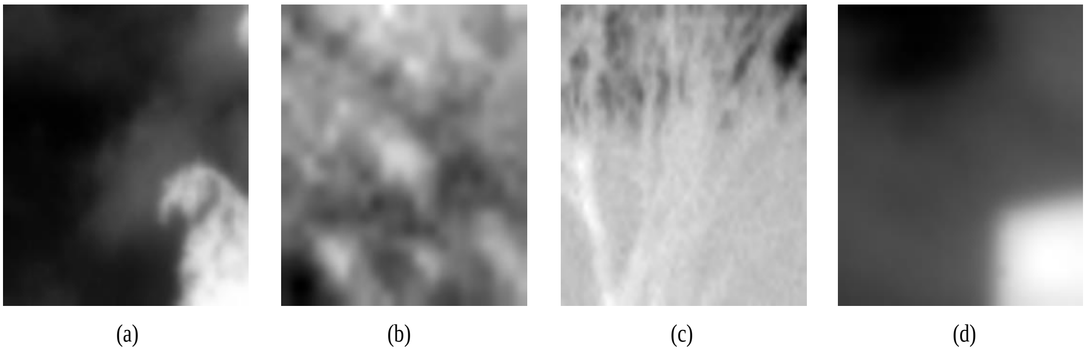


Figure 5.5: Typical false alarms are presented along with their sources. Beginning with (a) we can see the edge of a tree top, (b) shows the centre portion of a bush, (c) is a patch of muddy ground and (d) captures top corner of a building. These are illustrations of scenarios that tend to be detected and misclassified in ATR systems, especially in the LWIR band.

methods that we may require. In other words it has been shown to be a suitable dimension in previous works and offers *breathing room* for experimenting with network depth.

It follows that after enough data has been gathered to form a train/test set, each crop must undergo a resizing operation, as well as be mapped to $32pt$ precision where pixels exist in the range [01]. Lastly, a median filter with 3×3 kernel is convolved with each image crop. The intended effect is to remove trace spectral noise, as well as dead pixels, present in LWIR imagery. It should be noted that a consequence of spatial resolution adaption is slightly askew objects, such as the car shown in Figure 5.4 (b). This is due to the resizing of more rectangular images to square images. The data preprocessing step can be summarised and employed in the following order. The first step is to map the pixels from 14bit integers to single-precision. This is succeeded by applying a median filter. The last step is to resize the crop to 256×256. Once this preprocessing block is certain we have to evaluate the composition of the object imageset.

We can determine how many of each class contribute to the overall dataset structure by simply summing the total number of labels associated with each target. A balanced datset is one where instances are evenly distributed between classes and is preferred for training neural networks [166], purely because they have been shown to outperform networks trained on *imbalanced* sets. As shown in Figure 5.6 the initial dataset generated from collected Catherine MP sequences is very imbalanced, where half of the target classes have an overwhelming majority. If a network was trained

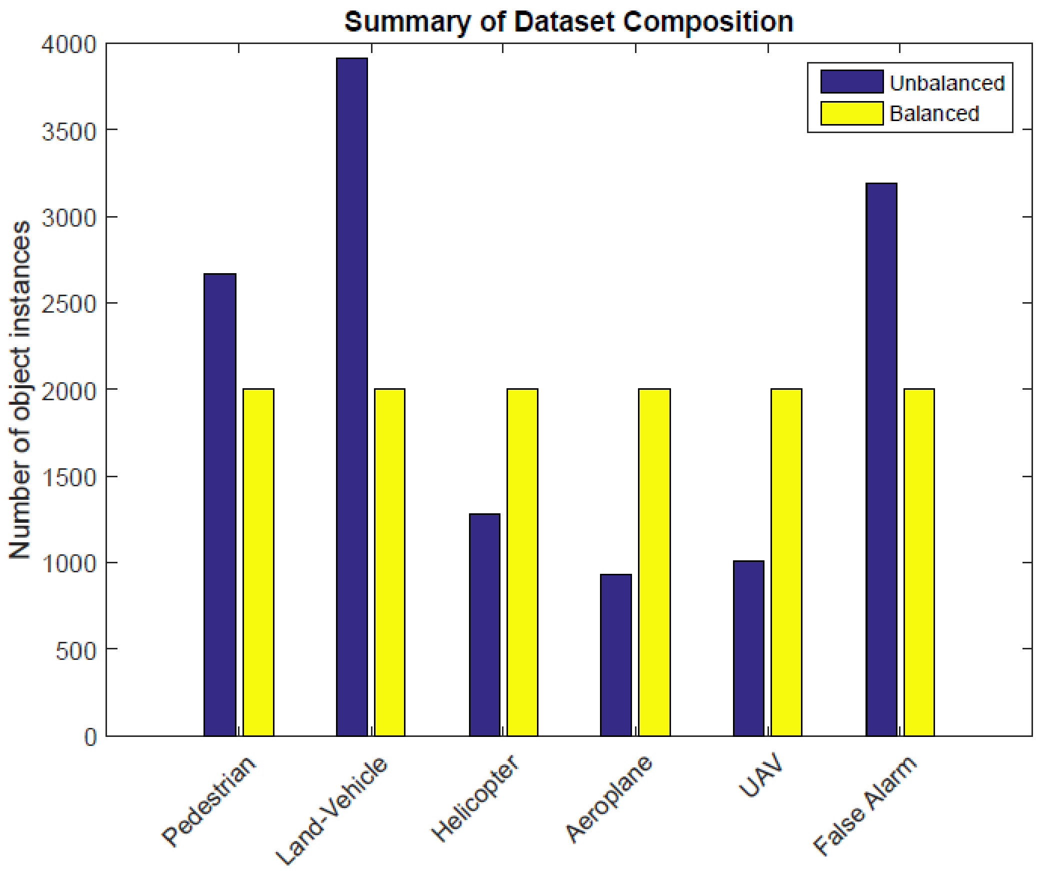


Figure 5.6: Bar graph showing the number of images present in the dataset.

using such a dataset, it would be biased towards the majority class and effectively *learn* that more images tend to be from these classes than others. This would lead to ineffective performance when deployed in real-world scenarios. Thus, it is beneficial in the long run to balance the dataset before embarking on a training scheme.

To address and re-balance the dataset two common approaches exist. The first is to randomly undersample from the majority classes, dropping selected images from the dataset. The second is to oversample from the minority classes by augmenting the existing data in some way to create new image instances, boosting examples per object class. To achieve these a desired number of examples per class has to be chosen, which we set at $\approx$ 2000 for each target class. The majority classes are undersampled accordingly via random exclusion until the set limit is reached. To augment data using existing imagery is slightly less non-trivial by contrast. For the minority classes, that is helicopter, UAV and aeroplane, random flips about the vertical image axis and whitened images are introduced. The image flips is simple to implement and whilst the overall pixel information is the same in flipped images, it appears at different locations from the source.

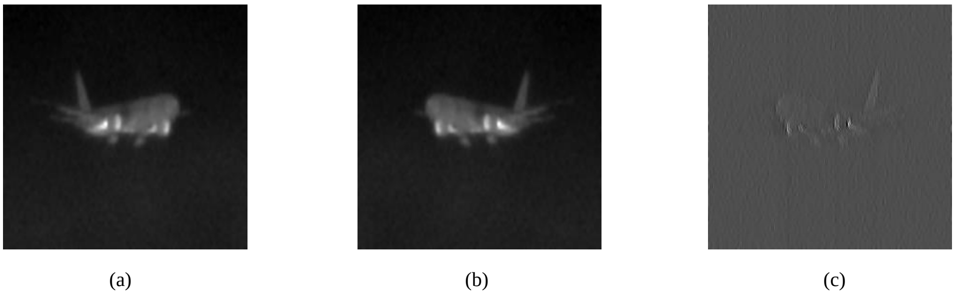


Figure 5.7: Image (a) is the original object instance of an aeroplane, cropped from a Catherine MP sequence. Image (b) is flipped around the vertical axis and (c) is augmented via whitening. The variations are added to the dataset.

Given most if not all objects in our dataset can exist at many different rotations or poses this seems a sensible option for data augmentation. The whitening of an image is a useful preprocessing step related to independent component analysis. The intended result is to transform an image matrix X into another image matrix Y, where Y has variance equal to unity and uncorrelated components [167]. The summary of image whitening is provided in Equation 5.1, showing the covariance matrix of Y equal to the identity matrix I.

$$E\left\{YY^T\right\} = I \tag{5.1}$$

Image whitening accentuates higher frequencies present in the image. In the case of LWIR source imagery it preserves hot-spots which may be crucial during the training phase. The effects of flipping images about the vertical axis and image whitening are illustrated in Figure 5.7. The minority classes are then randomly over-sampled and augmented using these two processes until the chosen instance limit is reached. After obtaining a balanced dataset as illustrated in Figure 5.6, we can design a CNN architecture and begin the optimisation process.

5.3.3 CNN Architecture

Supervised learning is the machine learning paradigm to which convolutional neural networks belong. As discussed in Section 5.2 CNNs have recently demonstrated to be very successful across a range of tasks. Their success can be attributed to feature extraction over multiple layers of increasing abstraction, leading to a better power of generalisation. Furthermore, non-linearities in the form of activation functions are introduced throughout the network. The combination of non-linear weights and

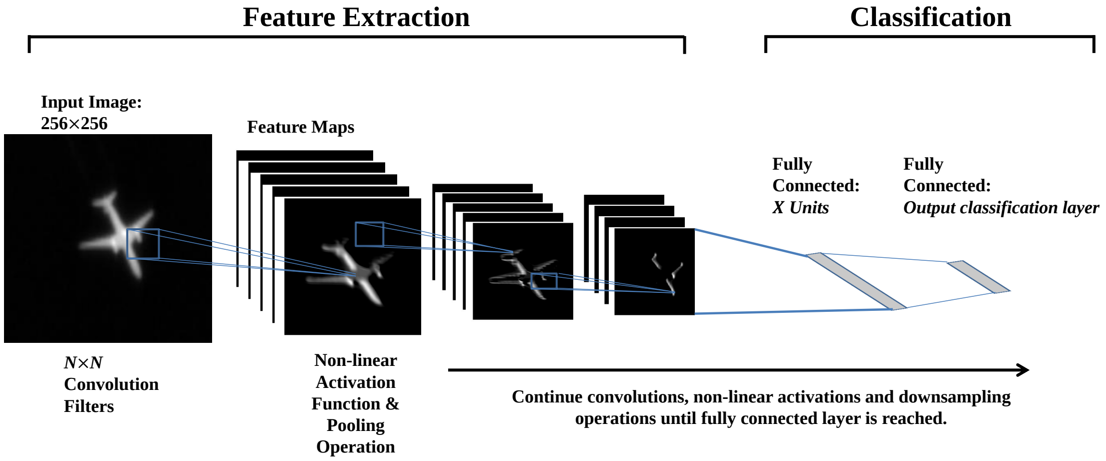


Figure 5.8: A typical CNN structure is presented, showing an example input image propagating through the network and generating feature maps. Each convolutional block is composed of a convolution, non-linearity and pooling layer. A deep network can be composed of many of these layers. Finally, the high level bstract features are fed into dense, fully connected layers which perform the classification task.

inputs determine a decision boundary capable of solving complex problems. The form of this boundary is discovered using a gradient-descent based method to minimise an objective function over the labelled dataset.

The general structure for a deep convnet is an architecture similar to those reported by Krizhevsky [41] and an overview of such an architecture is presented in Figure 5.8. The key component is the convolutional block composed of a tunable kernel (weights) for convolving input image arrays, producing *feature* or *activation* maps, a non-linear activation function in the form of a Rectified Linear Unit [168] (ReLU) $\phi(x) = max(0, x)$ and a downsampling pooling operation to reduce the spatial resolution.

Max pooling has been shown to be effective here. After an input image has propagated through such a network, with feature maps sufficiently extracted and downsampled, it will pass to a fully connected layer where the image structure is not preserved. A fully connected layer is akin to a traditional artificial neural network and is responsible for the classification aspect. It is common to see multiple fully connected layers at the end of a CNN structure, where the final output layer will be composed of units equal to the number of classes present in the dataset. Thus, in our case the final fully connected layer will contain 6 neurons.

The network structures as described in Krizhevsky [41] are designed to learn large-scale recognition tasks, such as the 1000 object class ImageNet dataset. Sufficient depth and width is required by a deep network to handle the complexity of such a task, where the number of parameters present in the network can grow remarkably large with increasing convolutional layers. For our problem domain we are only dealing with a 6 object class problem, using a dataset with only 12000 examples. Given the smaller nature of this recognition task we will not need to design a network with such a large structure. It should also be noted that the convolutional layers not only preserve the image structure and help generate invariant features, but they also keep the parameter space from growing exponentially. This is due to the smaller number of parameters needed in a convolutional block compared to a fully connected layer, which traditionally would have been used to tackle image recognition tasks where a huge parameter space becomes a limiting factor.

Although we do not require as deep a network as those used by Krizhevsky we still employ a symmetric network architecture with ReLUs and max pooling. Input images are of size 256×256 and small convolution kernels of 5×5 are also employed. We elect for a max pooling operation of size 2×2. Convolutions and pooling are implemented with stride lengths of 1 and 2 respectively. Two fully connected layers are always present at the end of the CNN. Although we shall explore the effect of depth and network width, the size and stride length of convolutional and pooling kernels remain constant.

5.3.4 Training Strategy

The assembled dataset, discussed in Section 5.3.1 contains 12000 LWIR object instances sampled over the 6 object classes, containing 5 real targets and a false alarm class. Training and test sets are created by splitting the dataset using a 90 : 10 ratio, where the test set is composed of 200 examples from each object class. Thus the training and test set total 10800 and 1200 examples respectively. The tunable network parameters are not affected by examples in the test set, remaining effectively *unseen* during the training phase. Network architectures are defined before each training run we and perform the training phase using opensource deep learning frameworks. We further test the ATR capabilities of CNNs by deploying a trained network in realtime on accelerated hardware.

Layer weights are initialised using the Glorot scheme [169] and every convolutional layer is composed of a number of filters, where we alter this value across layers as an experiment to explore network width. Convolutional blocks are followed by non-linear activation functions and pooling as discussed previously. The last convolutional stage feeds into a dense, fully-connected layer with a varying number of units, where we apply 50% dropout [144] to layer inputs. The random dropout of connections has been shown to improve network performance by stopping the CNN becoming reliant on certain neurons during training, increasing the generalisation power. Again, the ReLU operation is applied here. The final network output layer feeds into a softmax function [170], also known as a normalised exponential operation, with a fixed value of 6 units. The softmax outputs a pseudo-probability distribution over all classes which sum to one, where the probabilistic interpretation of this classifier function is:

$$P(y_i|x_i) = \frac{e^{f_{y_i}}}{\sum_j e^{f_j}} \tag{5.2}$$

This indicates the a normalised probability value assigned to each label y for an input image x, where f is an output score vector.

Training a neural network is a non-convex optimisation problem to minimise an objective function, where internal network weights are incrementally altered until a chosen error measure is diminished. We employ categorical cross-entropy loss as the objective function which is the common choice for multi-class problems and softmax outputs. Furthermore, backpropagation is employed in conjunction with the gradient-descent optimiser Adagrad [171]. Backpropagation allows the calculation of gradients for the cross-entropy loss function, with respect to global network weights. The computed gradient is then used by the gradient-descent algorithm Adagrad to allow network weight updates, where the goal is to minimise the overall error or loss function. The use of backpropagation is a standard training strategy for deep networks as it allows the network to see how changes to weights will affect the overall internal error.

Let us suppose that our objective function is $F(\theta)$, where θ is the model parameters. Gradient descent minimises $F(\theta)$ by updating parameters in the *opposite* direction to the gradient, given by $\nabla_\theta F(\theta)$. There is also an associated learning rate η responsible for the step size when descending the gradient slope. The learning rate is responsible for how effective the network learns. Too big a value for η and a network may never find a minima, whereas too small a value can exponentially increase the training time required.

To avert the tricky business of manually tuning and updating this hyperparameter η, we elect to use Adagrad. This gradient descent optimisation scheme updates the learning rate adaptively with respect to the parameters θ. The adaptive ability allows a large initial learning rate to be set, which can then be reduced accordingly to find minima, speeding up the training process considerably. If we observe $d_{t,i} = \nabla_\theta F(\theta_i)$ as the gradient of the objective function for parameter θ_i at time step t, the update rule for each parameter at time t can be given by $\theta_{t+1,i} = \theta_{t,i} - \eta \cdot d_{t,i}$. Finally, the Adagrad update rule is computed using Equation 5.3 :

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{D_{t,ii} + \epsilon}} \cdot d_{t,i} \tag{5.3}$$

which effectively shows us how the algorithm modifies the learning rate η, for θ_i at time t, using previous gradient calculations. The term $D_{t,ii}$ is a diagonal matrix where the main diagonal elements i, i are the sum of squares of the gradients and ϵ is a small smoothing variable introduced to avoid division by zero. It should also be noted that while we use gradient descent optimisation it is not true stochastic gradient descent, which computes the gradient using a single sample.

We do however use minibatch gradient descent, where the batch size is fixed at 20 examples. This is more computationally efficient and helps reduce the time to converge on a suitable minima. Lastly, to further prevent overfitting, along with the dropout technique, we introduce a regularisation penalty in the form of the ℓ_2 norm [172]. This ℓ_2 penalty discourages large weights in the network and is given by Equation 5.4:

$$\ell_2 = \sum_j \sum_k W_{j,k}^2 \tag{5.4}$$

The formula indicates that ℓ_2 is only a function of the internal weights of tensor W, the elements of which are squared then summed to give ℓ_2. This value extends the categorical cross-entropy loss L_i, which is a function of the data, after being multiplied by a small value λ, such that total regularised loss is given as $L_{reg} = L_i + \lambda\ell_2$. With access to a balanced dataset and a suitable optimisation schema, we can now explore the suitability of CNNs for LWIR target classification.

5.4 Experiments & Results

In this section we investigate several aspects of training a CNN towards the task of LWIR target recognition for ATR applications. The algorithm is implemented on a machine using an NVIDIA TitanX GPU with Kepler micro-architecture. The impact of CNN architecture on the overall network accuracy is explored and sufficient evidence of a *well-tuned* network is presented using a variety of tools, such as confusion matrices and feature visualisation. Categorical cross-entropy loss is calculated for training and validation sets. This is computed between targets and predictions using Equation 5.5:

$$L_i = -\sum_j t_{i,j} Log(P_{i,j}) \tag{5.5}$$

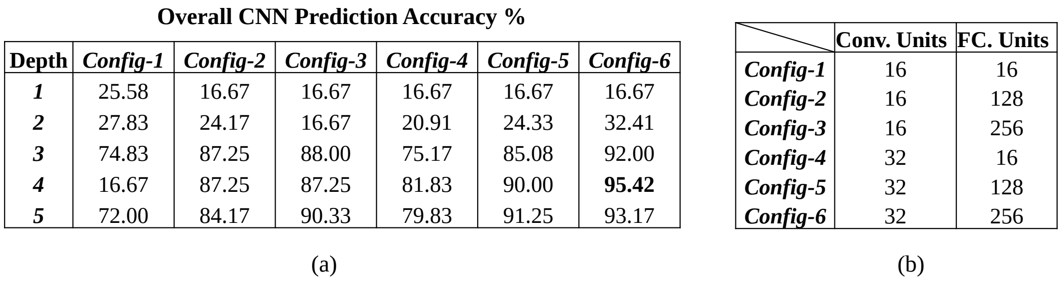
where $P_{i,j}$ is the softmax output for predicting classification probability and $t_{i,j}$ is the true target label. This loss function has a basis in information theory, but it can be interpreted as a kind of measure of internal network error. Target predictions on the validation set provide an insight into the networks ability to generalise and performance over this set of imagery is reported. This is achieved by comparing CNN predictions with actual class labels, which we construct into a confusion matrix, cfm. Overall prediction accuracy is given as $Acc = \frac{tr(cfm)}{N}$, which the trace, $tr()$, of the confusion matrix and dividing by total number of examples N in the validation set.

Given this is one of the first attempts to construct a CNN for LWIR target classification, over a variety of different object classes, using bespoke high-quality thermal imagery, there is no existing dataset or method to reliably compare our results against. Instead, the aim of this work is to convey and convince the community that recent deep learning methods are just as applicable to recognition applications in the thermal domain as those presented for colour images. We begin our investigation by training with shallow network structures and gradually increase their complexity.

5.4.1 Increasing Network Depth & Width

We first investigate the effect of varying CNN width, depth and number of units in the fully connected layer prior to the final output layer. The input dimensions, ReLU and pooling operations remain in place as described in Figure 5.8. We choose 6 network configurations for each layer / convolutional block, where the number of convolutional filters can be 16 or 32. For each of these we also vary the number of units in the first fully connected layer prior to the output, where we choose values of $16, 128$ and 256. This represents 6 variations per convolutional block, where a convolutional block is defined to be a $ConvolutionalLayer \rightarrow ReLu \rightarrow Pool$.

Each convolutional block has a depth of 1, so the stacking of convolutional blocks increases the overall *depth* of the network. Regardless of network depth, each block will have the same structure as defined by the configuration for number of convo-



Overall CNN Prediction Accuracy %

Depth	*Config-1*	*Config-2*	*Config-3*	*Config-4*	*Config-5*	*Config-6*
1	25.58	16.67	16.67	16.67	16.67	16.67
2	27.83	24.17	16.67	20.91	24.33	32.41
3	74.83	87.25	88.00	75.17	85.08	92.00
4	16.67	87.25	87.25	81.83	90.00	**95.42**
5	72.00	84.17	90.33	79.83	91.25	93.17

(a)

	Conv. Units	FC. Units
Config-1	16	16
Config-2	16	128
Config-3	16	256
Config-4	32	16
Config-5	32	128
Config-6	32	256

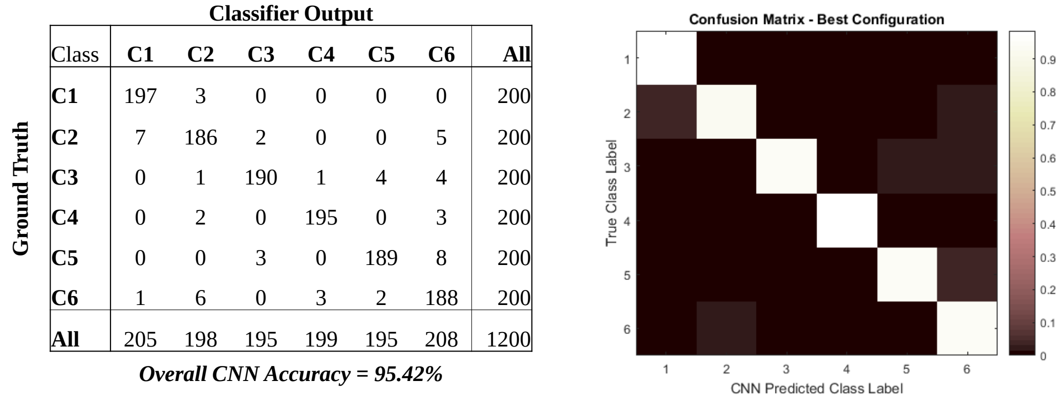
(b)

Figure 5.9: Table (a) shows CNN prediction accuracy, as a percentage, using the unseen test set for every network configuration. The best accuracy and configuration is highlighted in bold, identifying a depth of 4 convolutional blocks, each with 32 filters and a fully connected layer containing 256 units as the best performing CNN. Table (b) is a key to the configuration details, describing how many convolutional filters are in each block and the total number of neurons in the fully connected layer.

lutional filters and units in the fully connected layer. The variation experiment is summarised fully in Figure 5.9, where overall accuracy of the CNN over the test set is presented, for increasing depth (stacking convolutional blocks) using the described block configurations.

This experiment provides a granular insight into the effect of network structure on prediction performance, helping to identify which configuration is best to use. As such, we find that a CNN containing 4 convolutional blocks, each composed of 32 convolutional filters, ReLU and pooling layers, feeding into a fully connected layer of 256 units before the final output prediction layer, is best suited for our task at hand. This architecture paired with the outlined training scheme gained an overall prediction accuracy of 95.42%. A closer inspection of this networks predictions versus the true labels is given in Figure 5.10.

Furthermore, we can also plot how overall prediction accuracy increases as the training/validation decreases over time. In other words the network improves its ability to recognise LWIR objects as it learns from an increasing number of examples. Such a graph documenting the decreasing loss is presented in Figure 5.11 and it is a useful tool to identify if a CNN is overfitting. By observing how training and validation loss evolves through learning, we can see that the losses closely follow each other. They decline and plateau at roughly the same time which is a sign of a good fitting network. The validation accuracy, plotted against the second Y-axis, correctly increases as the loss is reduced.



Classifier Output

Ground Truth

Class	C1	C2	C3	C4	C5	C6	All
C1	197	3	0	0	0	0	200
C2	7	186	2	0	0	5	200
C3	0	1	190	1	4	4	200
C4	0	2	0	195	0	3	200
C5	0	0	3	0	189	8	200
C6	1	6	0	3	2	188	200
All	205	198	195	199	195	208	1200

Overall CNN Accuracy = 95.42%

(a)

(b)

Figure 5.10: Image (a) is the confusion matrix for our best performing CNN, showing the predicted output versus the true label for each example in the unseen test set. The main diagonal is ideally where the largest numbers should occur, as this would indicate the predictions match the true label. A graphical illustration of this confusion matrix is presented in image (b) as a heatmap. It is the same information but perhaps more intuitive to comprehend.

A sign of overfitting would be if the training and validation losses diverge or converge. Note that the significant gap between training loss shown as the red line and the validation shown as blue, is due to the additional ℓ_2 loss included in the training scheme to prevent overfitting. Again, this plot is generated using the ideal network architecture determined earlier. Having explored the performance of our CNN for LWIR target recognition in terms of prediction accuracy, we can also investigate the internal network structure to further clarify a deep convnets suitability to the task.

5.4.2 Network Visualisation

A common strategy when training a CNN is to visualise the inner mechanics of the network after convergence on a minima and optimal parameters. One intriguing area to observe is the filters in the first convolutional layer, as they can indicate a well-tuned network. It is very common to see the first layer filters from the work of Krizhevsky on ImageNet [41] presented in related publications. This is due to the filters being very distinct, indicating a finely tuned network, but also because many other works use these low-level features in their own networks as a starting point for further training. These iconic filter weights, along with our first layer weights and weights from a CNN trained over the MNIST dataset [173] are presented in Figure 5.12. As we can see from this image there is quite a large discrepancy between the highly orthogonal filters of the network tuned over ImageNet and the networks that

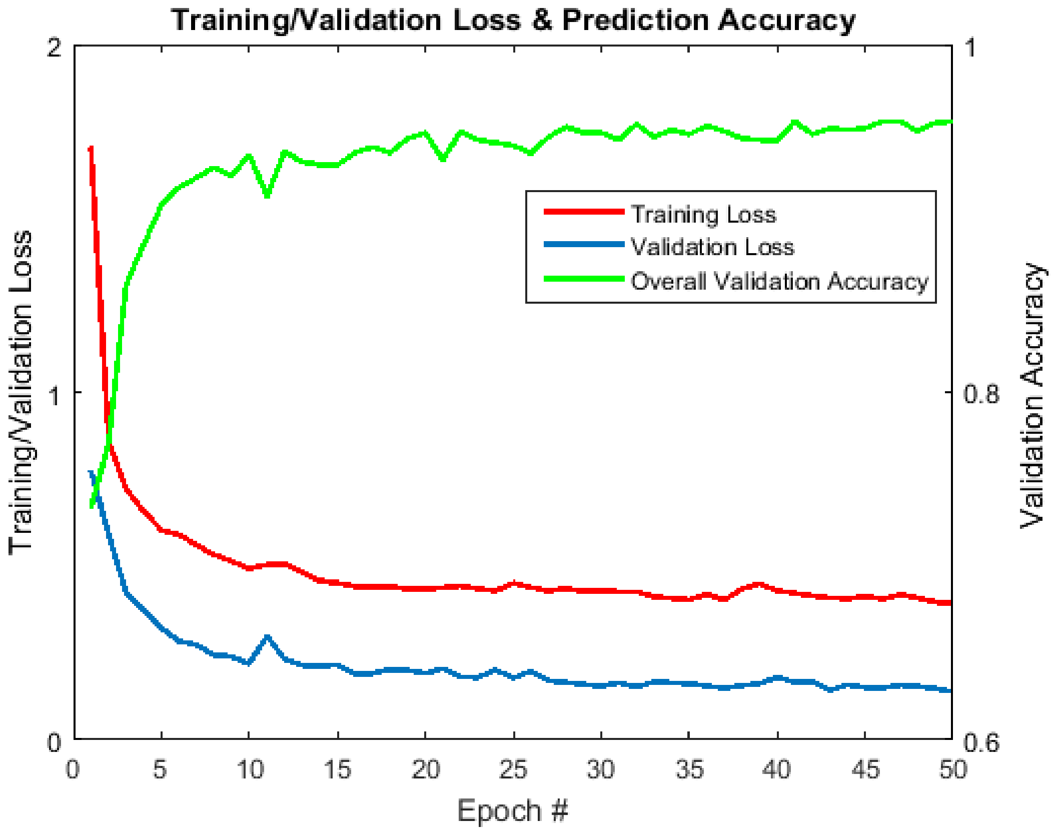


Figure 5.11: The graph is showing three things. As the learning process evolves, shown as an increasing epoch number, the training and validation loss is reduced. The validation accuracy, plotted against the right hand side Y-axis, quite correctly increases in correspondence with the decreasing loss. Lastly, the graph suggests the trained network has not overfitted to the data.

are not. However, this difference does not indicate that our best performing network is not trained well.

We can also examine the effects the tuned convolution filters actually have on an input image. Using the 32 learned filters as shown in image (b) in Figure 5.12, we can visualise the resulting filter response of the image as it is propagated through the first convolution layer. Each convolution kernel generates an activation map and we present an example using our tuned first layer weights for an input image of a pedestrian in Figure 5.13.

The main benefit to carrying out this technique is that it provides an insight into how the network is affecting input imagery, which can tell us some interesting things about the training. In the example provided it should be observed that each of the 32 responses contain some information, even if it may be sparse at the first layer. If we were to observe a few or more *dead* activations however, where maps

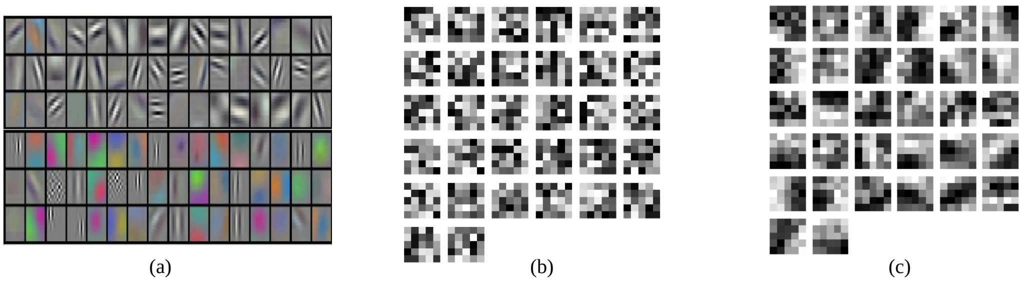


Figure 5.12: Image (a) shows the first layer weights from Krizhevsky's well known work on the ImageNet dataset. Image (b) is the first layer weights from our 32, 5×5, tuned convolutional kernels. Image (c) also shows 32, 5×5, tuned convolutional kernels but they have been trained over the MNIST digit recognition dataset, where the network achieves a prediction accuracy $> 99\%$.

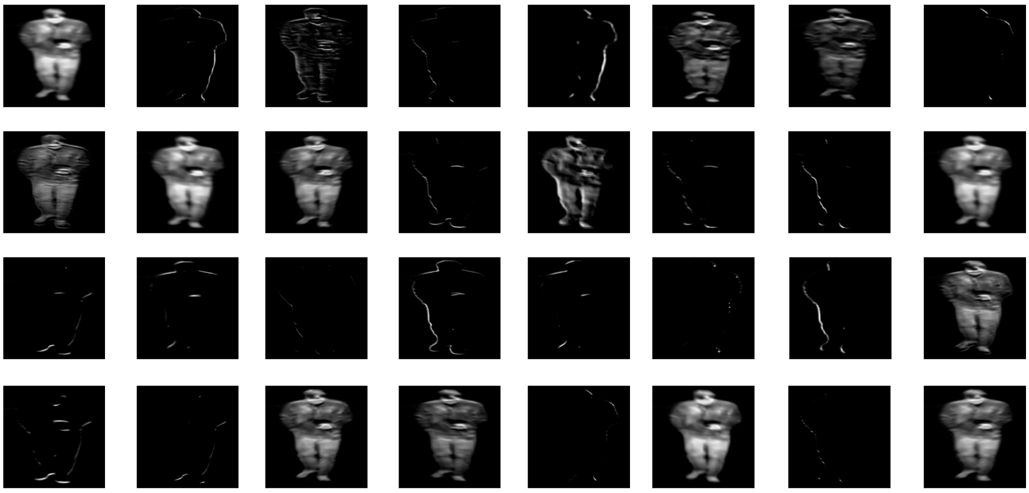


Figure 5.13: We illustrate the feature maps produced by each convolution kernel during a forward pass of an input image through the network. These are generated at the first convolutional layer and show that at this early stage the activations can remain fairly dense.

contain no information at all, it could signal a problem with the training scheme in place, such as high learning rates.

Lastly, we can exploit another tool to examine how well the trained CNN is at distinguishing classes using the validation set in a different way. Borrowing a similar idea as shown in the work of Mukherjee [174], we can use Linear Discriminant Analysis (LDA) to calculate a linear function of the output classes and plot the resulting measurements. In other words we are going to create an LDA projection of the validation set to observe some differentiation between classes, instead of feeding it through a softmax classifier. It is just another tool that helps verify the CNN is behaving as required.

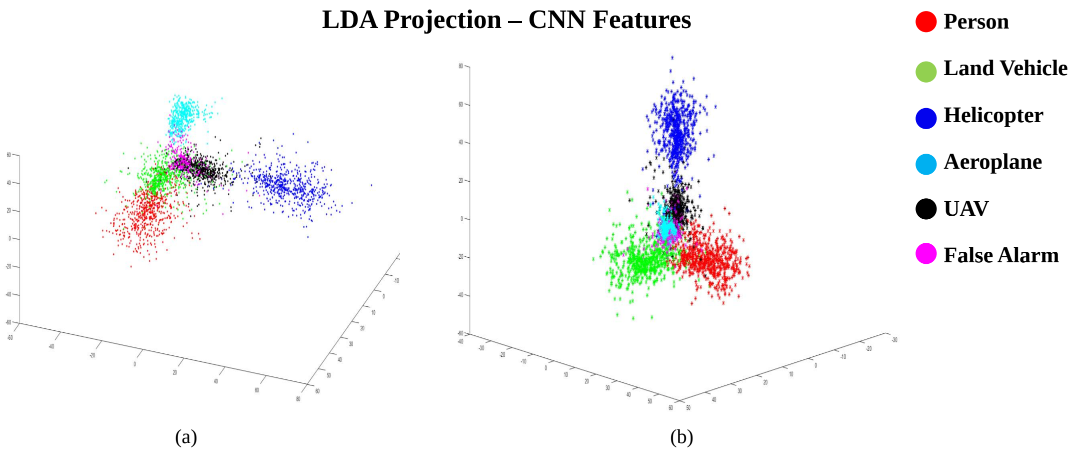


Figure 5.14: The LDA projection using CNN features from our trained network is shown from two different viewpoints in image (a) and (b). The visual key to identify classes is also shown on the right hand side. The plot shows that the clusters are mainly distinct and well separated.

The generated scatter plots are presented in Figure 5.14 and show two different viewpoints for the linear projection. This is necessary as we can only plot in 3-dimensions, whereas LDA produces as many linear functions as there are classes, so the rotated and angled view helps reaffirm the classes are distinct. In our case LDA is maximising class discrimination using 6 linear functions. This is the last method we employ to gain some insight into internal network performance. We follow up on this by taking the best performing network architecture identified in Section 5.4.1 and retrain a CNN using the same data, but in a different deep learning framework.

5.4.3 CNN Performance Verification

The final investigation we undertake is the retraining and deployment of our LWIR CNN in realtime, using the Caffe framework [175]. We simply define the best performing network architecture and retrain a CNN using the same dataset we created earlier, but using the C++ based Caffe environment. Once a comparably performing CNN model is obtained further verification experiments can be approached. The advantage of deploying a C++ based Caffe CNN is the ability to embed the network in a realtime system of our own design. Vast quantities of unseen Catherine MP footage can then be fed through an OpenCV framework where the video sequence is split by a frame grabber, allowing selected Regions of Interest (ROI) to be passed to the CNN. The network then observes and provides classification results. In other words we don't have to waste time carefully assembling a new LWIR object dataset for more testing, but instead can pass in a stream of images and simply draw boxes round targets of interest. It presents a final test to examine whether deep convnets truly are suitable for ATR applications in the thermal domain.

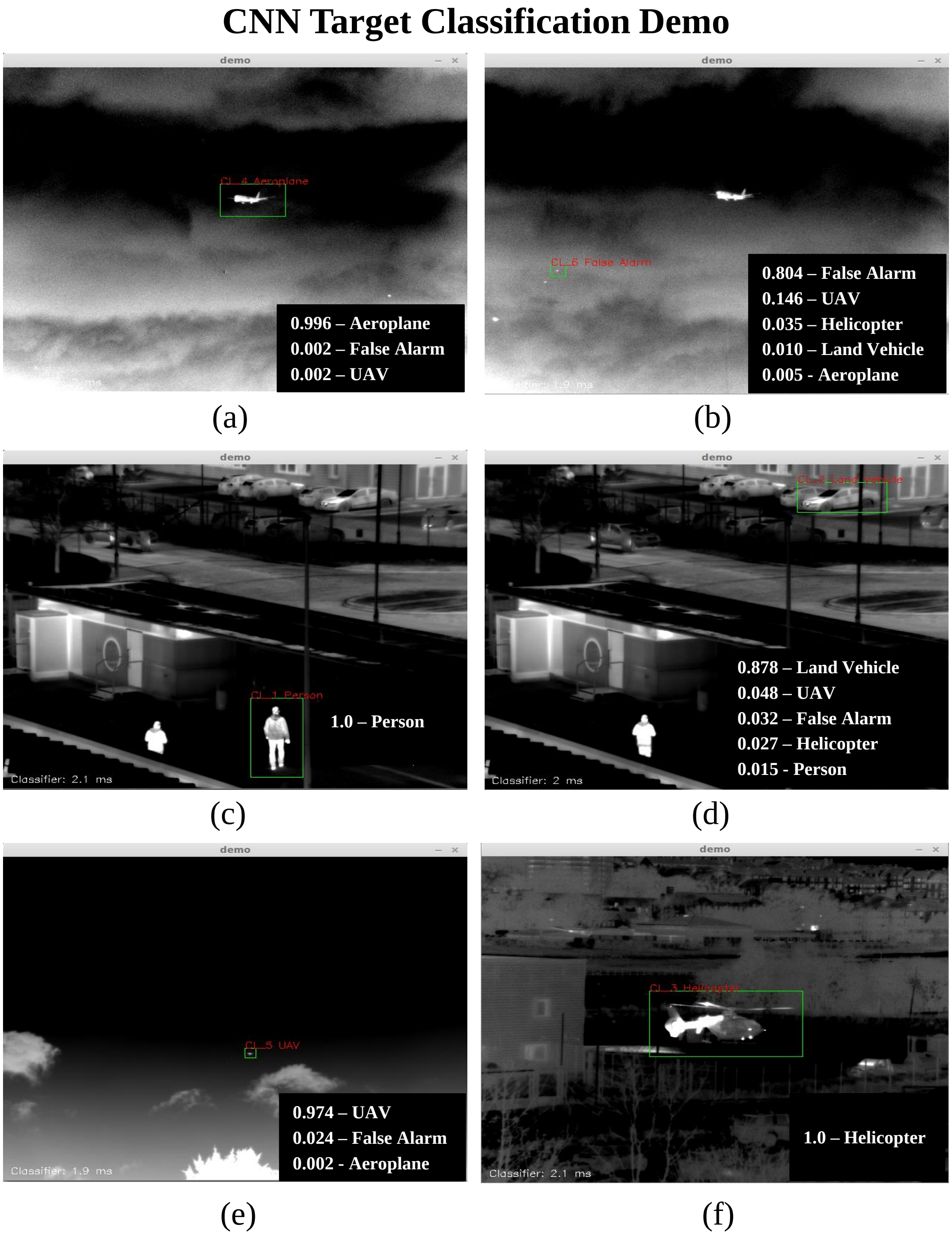
Having retrained the CNN model in Caffe using the same set-up as the best performing framework, the overall prediction accuracies between the deep learning frameworks are determined to be within 1% of each other, which could be down to small differences in initialisation etc. Once this comparable model is obtained we provide many examples to the network for classification and display the class scores. Examples of this final validation phase are illustrated in Figure 5.15 where a target classification for each object class is presented.

We deploy a CNN model using the best performing framework and optimal weights, so we may provide examples to the network for classification and display the class scores. Examples of this final validation phase are illustrated in Figure 5.15 where a target classification for each object class is presented. In this illustration, Image (a) is an aerial scene where the target is a passenger aircraft, image (b) is the same scene but the target this time is a passing bird, image(c) illustrates a typical urban surveillance environment where the target is a pedestrian, image (d) is the same scene where the target is a parked car, image (e) shows another aerial scene but the target is a small UAV and image (f) presents a helipad with the target being a grounded helicopter.

Thus far we have created a balanced dataset, determined the best CNN architecture to use, presented the prediction accuracy of this network configuration and visualised its inner workings in the form of convolution weights and filter response maps. Furthermore, we have demonstrated an LDA projection using the CNN features, illustrating the discriminatory power of deep CNNs towards recognition tasks. Lastly, we reimplemented an existing, high performing network in another framework and provided further examples of the classification capability for LWIR targets. This should be sufficient evidence of CNNs potential for ATR in the LWIR domain. As such, we can now discuss the findings of the presented experiments and results.

5.5 Discussion

Developing and demonstrating a deep convolutional network capable of classifying objects of interest in LWIR imagery, to a high degree of accuracy, was the grand


CNN Target Classification Demo

0.996 – Aeroplane
0.002 – False Alarm
0.002 – UAV

(a)

0.804 – False Alarm
0.146 – UAV
0.035 – Helicopter
0.010 – Land Vehicle
0.005 - Aeroplane

(b)

1.0 – Person

(c)

0.878 – Land Vehicle
0.048 – UAV
0.032 – False Alarm
0.027 – Helicopter
0.015 - Person

(d)

0.974 – UAV
0.024 – False Alarm
0.002 - Aeroplane

(e)

1.0 – Helicopter

(f)

Figure 5.15: This figure presents example classifications for each target present in the dataset, given as images (a - f). Objects passed to the CNN are identified in the image by the green ROI box enclosing the target. Pseudo-probability scores are provided in the lower right hand corner of each image, showing the output class predictions of the trained convnet.

aim of this work. Furthermore, we intended to show such a network not purely in terms of classification accuracy, but to delve into the training scheme and visualise the finer details of a trained CNN. The reasoning behind this was to present overwhelming evidence that modern deep learning methods were indeed suitable for use with thermal band imagery.

The motivating factor to design such an algorithm can be explained by the desire to harness the potential benefits for a robust recognition tool in ATR applications, especially in the security and defence domain which primarily relies on thermal imagery. Ultimately, we believe these goals have been achieved by the determination of our results, as well as highlighting some interesting insights along the way. The principal implications of our experimental findings can be summarised through a discussion of Sections 5.3 and 5.4.

Network performance and accuracy results are only presented in the form of confusion matrices. We build on this and explore the underlying machine learning aspect of training a CNN for LWIR target recognition, whilst also extending the understanding of the trained CNN by visualising fundamental network behaviour.

5.5.1 Data Cleansing

One of the first steps we undertake is to address the class imbalance present in the LWIR dataset as a network trained on an unevenly distributed set of images can lead to biased performance. If we look at Figure 5.6, the number of instances present for each class shows a heavy bias towards pedestrians, land vehicles and false alarms. Thus, a network trained on this will tend to be better at predicting these examples. Furthermore, a validation set randomly sampled from an imbalanced dataset will also be unevenly balanced, favouring the majority classes.

So, the biased network will also appear to be display accurate performance and it may not be easy to determine if this is a detriment without significant effort. This imbalance and bias network / validation set is immediately resolved via augmentation and undersampling. Once the dataset is balanced we can create even distributions of class in both train and test sets, ensuring reported performance is indicative of true network performance when deployed and faced with unseen imagery.

5.5.2 CNN Optimisation

The training strategy and general network structure we employ, as shown in Figure 5.8, is typical for creating a robust CNN towards recognition tasks. Our contribution in this area is the exploration of network width and depth affecting predictive performance. The network architectures were designed to gradually become more complex and overall classification accuracies were recorded using the same validation set. If we observe Figure 5.9 it summarises the different network configurations used.

Curiously, for some of the smallest architectures it appears the overall accuracy is better than deeper networks by comparison. For instance, configuration 1 at depths 1&2, the accuracies are higher compared to the deeper nets with the same width etc. The observed volatility is actually due to the stochastic nature of training a neural network, where initial parameters are set by a random process. It would have been better to repeat this experiment multiple times and average the accuracy results, eliminating the stochasticity factor. However, the experiment still allows a general trend to determined. Perhaps unsurprisingly we can see that a deeper network with reasonable width achieves a greater accuracy on average, with the top performing structure containing 4 convolutional layers and a denser fully connected layer.

This corresponds well with the current literature which suggest compact convolutional kernels and deeper network structures perform better than shallower nets. The underlying reasoning behind this observation is that deeper nets have a better chance of recovering more abstract and generalisable features present in the imagery than shallower nets would, which leads to better classification accuracy. The confusion matrix of our best performing network structure shown in Figure 5.10 confirms that the trained CNN performs well across all classes, resulting in the peak overall accuracy.

5.5.3 Analysing CNN Visualisation

Once this finely tuned CNN is obtained an additional avenue is explored through the visualisation of first layer convolution filters and activation maps, shown in Figure 5.12 and Figure 5.13 respectively. The filter response for a target pedestrian propagating though the first layer exhibits ideal behaviour for a well trained CNN. Each of the activations presented in Figure 5.13 contains some information, with many being dense with pixels. This is a good sign as empty or *dead* filters would indicate a poor training regime.

The most interesting insight discovered though arises when we consider the convolution kernels shown in Figure 5.12. A great many related works present the nicely tuned, orthogonal filters from a CNN tuned using the ImageNet dataset as shown in Figure 5.12(a). However, when we visualise the filters from our finely tuned network they appear noisy and with little structure. This is reportedly a sign of a badly trained network which is of course worrying. Yet, we explore this issue further by plotting the first layer filters from a network trained over the MNist character dataset, where the network practically has perfect recognition. These filters are shown in Figure 5.12(c) and surprisingly also exhibit noisy filters, despite the network exhibiting a near flawless accuracy.

This suggests that the low-level features obtained for a high performance CNN do not always need to be *clean* and show strong orthogonality. In other words noisy filters are not indicative of a poorly tuned or inaccurate network. We propose that the small number of classes and subsequent little variation in the imagery contained within our LWIR object dataset, as well as the MNist data, gives rise to noisier filters. The reasoning is that CNNs simply do not have to learn such clean, discriminative low-level features when the problem is small in scale and constrained.

By contrast the networks trained over ImageNet will eventually have to converge on such nice filters as they are essential to differentiate between 1000 object classes with any degree of accuracy. To the best of our knowledge this interesting insight has not been communicated before, probably because most practitioners of CNN training for recognition tasks tend to use the well-tuned, low-level features as a starting point for their specific problem. We did not see much point in using these low-level features as thermal imagery is quite different to colour, so we elected to design and train CNN from scratch.

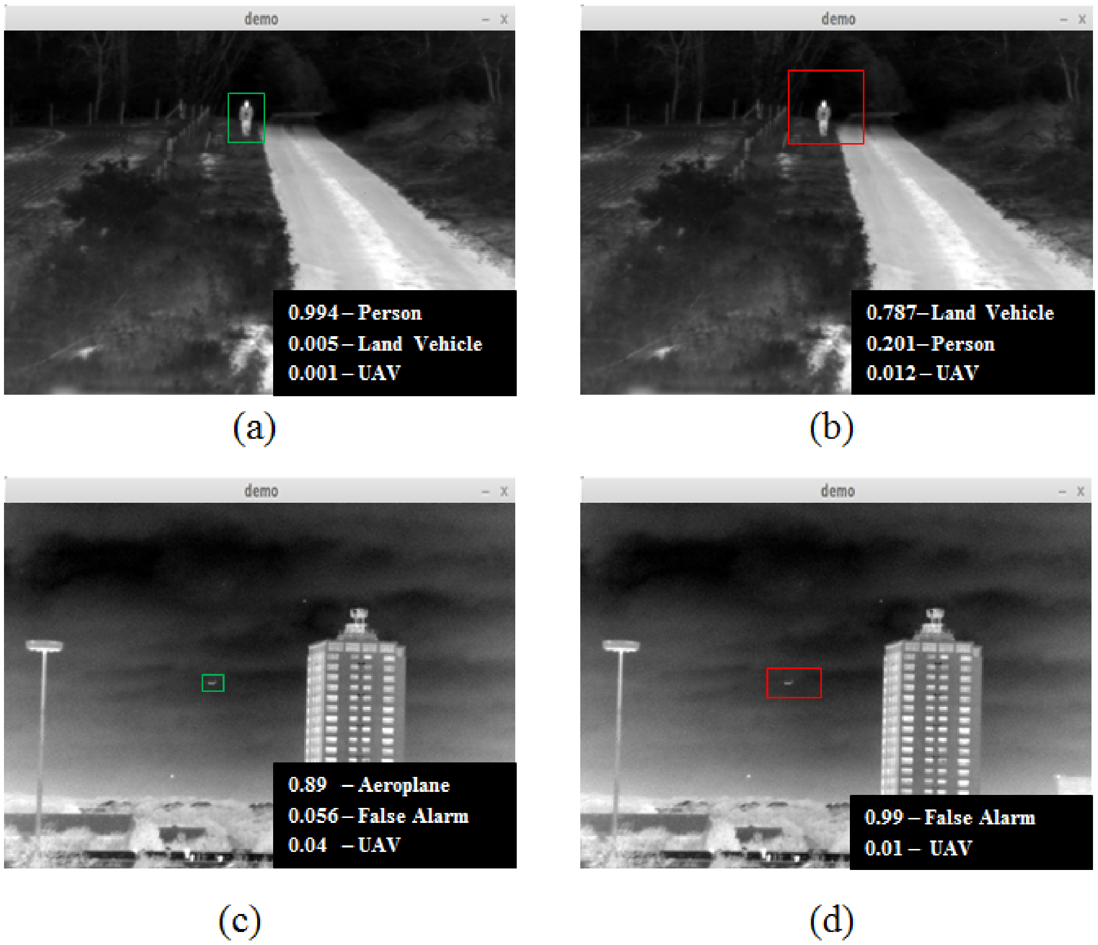
The assertion that our noisy convolution filters are no indication of poor training is further complemented by the evidence shown in Figures 5.11 & 5.14. The plot of training and validation loss decreasing as prediction accuracy increases is ideal behaviour when training a deepnet. The gap between loss values is due to the additional ℓ_2 penalty added into the training loss, so it is always higher, but they follow each other closely otherwise. A poor optimisation scheme would generate a much different looking graph. Furthermore, the LDA projection scatter plot illustrated in Figure 5.14 also demonstrates the learned CNN features are excellent at discriminating between target classes. This strengthens not only belief in our networks suitability for the chosen ATR task using LWIR imagery, but it also lends further credence to the notion that noisy low-level features are not indicative of a well-tuned and accurate network overall.

5.5.4 Realtime Deployment & Failure Modes

Our final investigation to verify CNN performance and potential inclusion into an ATR system using LWIR imagery required an implementation of an executable CNN on hardware. To this end we implement the best performing network architecture discovered prior on a GPU. This framework offers the capability to deploy a trained convnet in a realtime (classification is ≈ 2ms per image) environment, allowing us the opportunity to throw much more test data at the network without time intensive image cropping efforts. Example target classifications using this system are provided in Figure 5.15 showing a variety of scenes and targets, along with correct class predictions.

Let us now examine Figures 5.15 (a), (b) and (e). Both images (a) and (b) are of the same aerial scene but image (a) shows the passenger aircraft as the ROI/target and corresponding, accurate classification. Image (b) then selects a passing bird as the target to pass to our trained CNN where it is correctly classified as a False Alarm. While we have to acknowledge the correct classifications in the described cases, the real deductive power of a fine-tuned CNN is displayed when we observe image (e). This example is of another aerial scene but the selected target this time is a small UAV, which looks very similar to the passing bird shown in image (b). Yet, in these cases the network quite confidently predicts the correct class.

One very useful feature of the realtime demo is the ability to draw ROIs to explore how the trained CNN handles different aspect ratios and object poses. This offers a valuable insight into the failure modes of the optimised CNN. Some examples of misclassified images are provided in Figure 5.16, highlighting how important the bounding box or aspect ratio is for a convnet. This illustration strongly suggests that misclassifications arise from bounding boxes not being tightly coupled to the target thermal signature. Both of the wrong predictions are a result of the ROI being too big or wide compared to the object, so aspect ratio is very important to the trained network. This also hints at the training set being composed of object instances where the crop is also fairly tight to the object boundaries. This kind of insight is much easier to observe with an interactive demonstrator and can inform how an end to end ATDR system should be designed, with respect to target acquisition.



0.994 – Person
0.005 – Land Vehicle
0.001 – UAV

(a)

0.787–Land Vehicle
0.201–Person
0.012 – UAV

(b)

0.89 – Aeroplane
0.056 – False Alarm
0.04 – UAV

(c)

0.99 – False Alarm
0.01 – UAV

(d)

Figure 5.16: This Figure presents two thermal image scenes that have been processed using the realtime demonstrator and an ROI selected to be classified using the Caffe model CNN. Image (a) shows a pedestrian walking in a rural scene with the correct classification. Image (b) is of the same scene but the larger ROI results in a misclassification. Image (c) shows a wide area scene with an aircraft in the centre distance. Again, the CNN correctly predicts the target class. Image (d) is of the same scene but the CNN fails to classify the target from the much wider ROI.

Taking account of all the experimental results and subsequent discussions, it should be very clear that recent machine learning methods are quite suitable for tackling our stated goal. Namely, CNNs are an ideal tool to design a robust target recognition system in the thermal domain, specifically for LWIR imagery. The stated prediction accuracies and network behaviour correspond well with the literature and confirm the validity of our arguments.

5.6 Conclusion

In this chapter we present a novel convolutional neural network and demonstrate its suitability for LWIR target recognition applications. An extensive comprehension of network performance and behaviour is gleaned via experimental validation, where we show a wealth of evidence to confirm our findings. The only thing lacking is a direct comparison to similar methods on a benchmark set, but this is very hard to achieve due to the lack of competing methods and public, labelled thermal object datasets.

Classification and visualisation results are discussed in the context of the closest related work, of which this is a natural extension. An interesting finding illustrated that low-level network features need not be highly orthogonal to obtain reliable performance. It does suggest, however, that a finely-tuned network displaying noisy filters is likely to be tackling a constrained, small scale problem and highlights an area of future work in the process. Ultimately, we demonstrate that these learning models are well suited for thermal image recognition tasks, capable of providing intelligent signal processing within the security and defence domain. We shall now proceed to our final data chapter where we embed this trained network in an end-to-end model for challenging ATR scenarios using real world data.

Chapter 6
An Enhanced ATR System

In the previous chapter we outlined a complete process to construct a high performance target classifier using CNNs and LWIR thermal imagery. We demonstrate accurate classification accuracies across all classes in the dataset. Now we can pull together all the discussed research components to deploy the trained CNN to more challenging, real world data representative of long range surveillance scenarios that are commonplace in the defence community.

Not only do we utilise the optimised CNN model for ATR purposes but we also draw upon key ideas discussed in the previous chapters of this thesis, thus returning us to the grand aim of the project that we set out to solve. The challenge in this chapter is to develop a more complete system capable of handling multi-modal video sequences and providing enhanced autonomous target classification performance. The imagery is demanding as targets exist at very long ranges with subsequently small thermal signatures. The model we develop deploys the trained CNN to provide initial classification results, which are then affected using region segmentation and spatial context information. Target classification scores for all model variations are also presented.

The work outlined in this chapter ties together all the previous strands of research we have explored so far, where we ultimately demonstrate an improved ATR system and enhanced situational awareness. The applied CNN research reproduced in this chapter was presented at SPIE Security & Defence, Electro-Optical and Infrared Systems; Technology and Applications, 2016 [16]. Furthermore, the spatial context work enhancing long range ATR is due to be presented at SSPD London, 2017.

6.1 Introduction

As we have stated repeatedly throughout this thesis, intelligent signal processing capabilities are highly desirable for surveillance tasks within the security and defence domain. Automatic Target Detection (ATD) and Automatic Target Recognition are two critical aspects of surveillance based applications. We have already addressed the ATR problem in isolation in Chapter 5, so we shall refrain from in-depth explanation regarding the classifier construction again. We do, however, incorporate the trained CNN LWIR classifier into a larger Automatic Target Detection & Recognition (ATDR) system. This scheme utilises sensor information gathered from a combination of visible and thermal-band sensors and is a real application of our CNN model.

6.2 Motivation

Employing multi-modal sensor platforms for security related applications is a widespread practice, where each imaging sensor modality is sensitive to a different waveband of the EM spectrum [150, 176]. Utilising additional spectral bands provides increased knowledge of surrounding environments. The motivating scenario outlined in Chapter 1 is the driving factor for the research explored in this chapter and the thesis as a whole.

Recollecting the scenario, a typical realisation would be a land reconnaissance vehicle equipped with a multi-sensory platform and a crew tasked to provide relevant intelligence for a target scene. Ultimately the goal is to improve the overall *situational awareness* through effective exploitation of sensor information. However, each additional information source increases the burden on a user/operator to quickly process the incoming image data and accurately report findings. The extra load placed on an operator in a stressful, potentially hazardous, environment could have disastrous consequences if a crucial detail is overlooked.

Effective ATDR methods become invaluable as they address this problem scenario by automating the signal processing and alleviate the bulk of the task from a human user. The automatic system could, for example, remove extraneous details leaving only salient regions of interest, or highlight the most important aspects in the surrounding scene prioritised by threat level [177]. Both examples illustrate the system presenting an operator with a vastly reduced information load, but with a greater perception of surroundings, requiring significant effectual automatic signal processing methods.

There is an existing body of prior work focused on creating such techniques which our work indirectly improves upon and advances the field [150, 178, 179]. We achieve this by capitalising on recent machine learning methods to create an object classifier for high quality thermal image data,. The output of the object classifier can be affected by an overarching contextual framework utilising colour information providing spatial context. Thus, we should explore prior knowledge in three key areas: the surge in machine learning using Convolutional Neural Networks, existing IR target classification schemes and the role of scene context. Only the role of scene context remains to be explored for we previously reviewed the other key areas in Chapter 5.

6.3 Scene Context

Constructing a contextual framework to harness additional scene information is quite prevalent in the field, where *context* is understood to be any useful additional information relating to scene perception for the task at hand [95]. For instance, certain prior works [180, 181] model the relationship between global scene structure and object properties contained within, where scene configuration and spatial layout can be used to aid object localisation tasks.

A further example of using learned scene structures to enhance object localisation is presented in [182], which exploits effective scene categorisation to determine what objects are most likely to be found. If we understand how objects relate to regions within a scene by observing pixel level information surrounding objects, improved object recognition performance can be achieved [183, 184].

In our case we aim to enhance the ATR process for long range targets using semantic region segmentation, which provides the context surrounding each detected target. This idea is realised in related work, where the likelihood of surrounding regions for each object is used to determine the overall probability of an object class to exist within the scene [185]. We build upon and extend this contextual framework, incorporating CNN class scores, to suit our task of effective ATR for long range targets using multi-modal sensor information.

Overall, effective ATDR performance in challenging surveillance scenarios is still a highly sought after capability. The use of multi-modal sensors is increasing which introduces a trade-off between increasing processing complexity and better scene perception. As we shall explore, a semantic region segmentation process and object priors may offer effective long range ATDR performance when incorporated with a robust CNN classifier. The whole system relies on multi-modal sensor information and exploits the strengths of each waveband. Now we can move onto describing the design of our ATDR model.

6.4 Long Range ATDR

An end-to-end ATDR system using multi-modal input data can be outlined in this section. Algorithms design will focus on real-world, *noisy* multi-modal data, collected from a static surveillance platform in rural environments as illustrated in Figure 6.4. The system design is reflected in the layout of this section and contains three key elements. Given an input data stream the first stage is the generation of candidate targets via an ATD algorithm, briefly described in Section 6.4.1. Candidate detections can be passed to a trained LWIR object classifier for ATR processing, outlined in Section 6.4.2.

Output probability scores for each detection will serve as input into an overarching contextual framework, explained in Section 6.5, utilising colour band information as well as prior scene and object knowledge to affect final class scores. The system will initially be developed using mid to long range sequences for optimisation before being fully evaluated over several, longer range surveillance sequences, where all sequences require manual ground-truthing. We aim to follow a datascience paradigm when developing our solution, where we will employ *training, validation and testing* stages. Initial algorithm training and validation stages using small representative datasets will help to solidify network structures and improve the general approach, where potential improvements can be incorporated into the system. The final testing stage will thoroughly evaluate the systems effectiveness for the described problem using a challenging, unseen dataset.

6.4.1 Autonomous Target Detection System

The central theme of this work is concerned with enhancing overall ATR performance via CNNs and context mobilisation, meaning the choice of ATD algorithm is

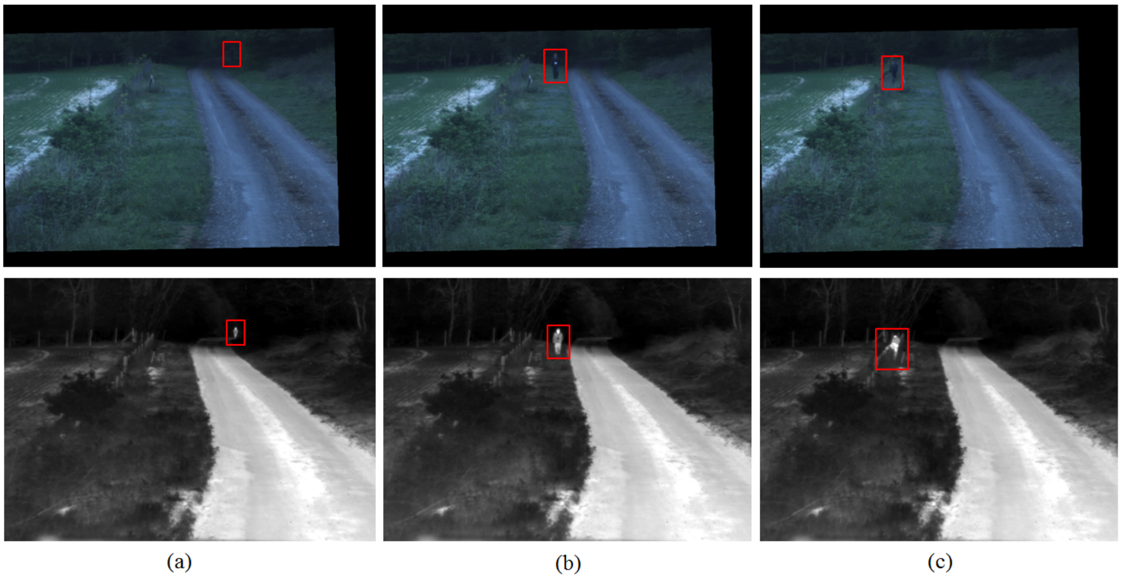


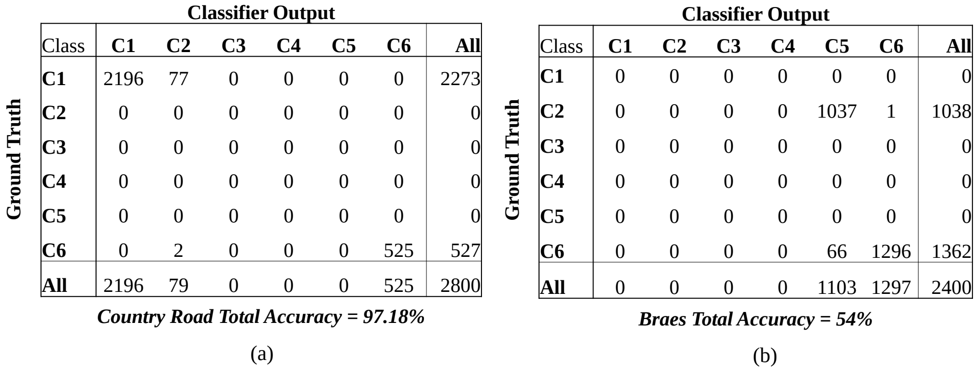
Figure 6.1: Short Range Sequence *Country Road*. The top row illustrates registered colour band imagery, showing a person walking toward the sensor platform along a rural path. The bottom row is the corresponding LWIR imagery showing the same scene. ATDR is performed only on the thermal data stream, but bounding boxes are shown on both modalities for clarity. [16]

of small significance as long as it can detect targets. Typically with thermal data some form of hotspot detection is employed to generate target regions, for example the method presented by Teutsch et al. [152].

Autonomous target detection is performed only on the thermal image feed from the Catherine MP LWIR. We use a proprietary Thales algorithm for this task, which is capable of localising targets from short to very long ranges. The crucial step is to ensure that the preprocessing steps used in training, such as those described in 5.3.1, remain the same in deployment. Target images need to be of the same dimension and format the CNN was trained on to ensure correct behaviour. The processed samples for each target are passed forward to the trained CNN for classification, along with localisation information for the context framework.

6.4.2 Initial Long Range Target Classification

Employing Thales' ATD algorithm on LWIR video sequences we can generate candidate targets to be classified with the trained CNN, giving an early-stage ATDR scheme. The first step for developing an effective end-to-end system is to explore ATDR process performance on some initial sequences, captured using the multi-sensor set-up shown in Figure 5.3 in chapter 5. The first scenario is a short to mid



Classifier Output

Ground Truth / Class	C1	C2	C3	C4	C5	C6	All
C1	2196	77	0	0	0	0	2273
C2	0	0	0	0	0	0	0
C3	0	0	0	0	0	0	0
C4	0	0	0	0	0	0	0
C5	0	0	0	0	0	0	0
C6	0	2	0	0	0	525	527
All	2196	79	0	0	0	525	2800

Country Road Total Accuracy = 97.18%

(a)

Classifier Output

Ground Truth / Class	C1	C2	C3	C4	C5	C6	All
C1	0	0	0	0	0	0	0
C2	0	0	0	0	1037	1	1038
C3	0	0	0	0	0	0	0
C4	0	0	0	0	0	0	0
C5	0	0	0	0	0	0	0
C6	0	0	0	0	66	1296	1362
All	0	0	0	0	1103	1297	2400

Braes Total Accuracy = 54%

(b)

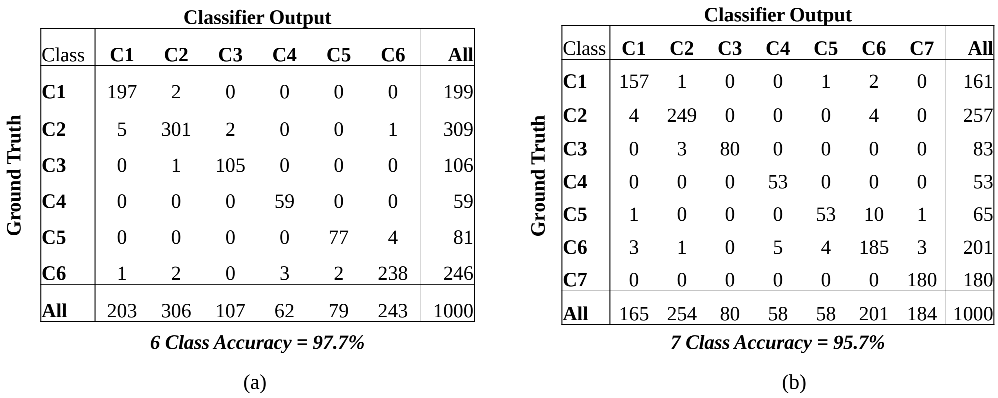
Figure 6.2: Exploratory classification results for two sequences. Subfigure (a) is output classification results for the *country road* sequence. The only classes present are **C1** & **C6** which are person and false alarm respectively. Subfigure (b) is output classification results for the *Braes* sequence. The only classes present are **C2** & **C6** which are land vehicle and false alarm respectively [16]

range clip in a rural environment, where a person walks towards the sensor platform as illustrated in Figure 6.1.

The overall evaluation results for the *Country road* sequence are presented in Figure 6.2, confirming the effectiveness of our CNN for LWIR target classification. This is perhaps unsurprising given the training imagery is of similar quality to the targets from the *Country road* sequence. The second scenario is a much longer range sequence, *The Braes*, which we again use to assess the early-stage ATDR scheme. The imagery from this sequence is illustrated in Figure 6.4.

As we can see in Figure 6.4, ATDR for this scenario is very challenging. The target is very small in resolution and low quality. We specuatively apply the initial ATDR scheme to this long range scenario and examine the classification results. Somewhat unsurprisingly the CNN is not capable of assigning correct class labels to targets of such low image quality, highlighted by the low accuracy and confusion matrix results presented in Figure 6.2. However, upon closer examination it appears the network can actually differentiate between false alarms and real targets with resounding accuracy, it just gets the object class consistently incorrect.

Given the network has not seen images of such low quality in the training process, we capitalise on this capability by gathering long range training example, similar to subfigure (c) in Figure 6.4, to allow retraining of the CNN model. The network is optimised with additional long range target imagery, where the final fully-connected



Classifier Output

Ground Truth

Class	C1	C2	C3	C4	C5	C6	All
C1	197	2	0	0	0	0	199
C2	5	301	2	0	0	1	309
C3	0	1	105	0	0	0	106
C4	0	0	0	59	0	0	59
C5	0	0	0	0	77	4	81
C6	1	2	0	3	2	238	246
All	203	306	107	62	79	243	1000

6 Class Accuracy = 97.7%

(a)

Classifier Output

Ground Truth

Class	C1	C2	C3	C4	C5	C6	C7	All
C1	157	1	0	0	1	2	0	161
C2	4	249	0	0	0	4	0	257
C3	0	3	80	0	0	0	0	83
C4	0	0	0	53	0	0	0	53
C5	1	0	0	0	53	10	1	65
C6	3	1	0	5	4	185	3	201
C7	0	0	0	0	0	0	180	180
All	165	254	80	58	58	201	184	1000

7 Class Accuracy = 95.7%

(b)

Figure 6.3: Validation results generated by trained CNN over 1000 unseen LWIR examples for 6 & 7 target classes. The overall CNN accuracy is 97.7% for 6 classes, which drops slightly to 95.7% with the presence of an additional 7th class. Ground truth classes are rows, classifier outputs are columns. The main diagonal reveals classifier performance. **C1** is the person class, **C2** is land vehicle, **C3** is helicopter, **C4** is aeroplane, **C5** is UAV and **C6** is the false alarm class. The additional $7th$ object **C7** is the long range target class [16] .

layer outputs over 7 units instead of 6 to account for the additional class. Output validation results from retraining the LWIR CNN are presented in Figure 6.3, subfigure (b).

The successful discrimination of real targets and false alarms is achievable via the introduction of a long range target class, determined by the exploratory training and validation stages. However, we still do not know the actual object class of real targets but only that it is an object of interest. The key benefit of introducing a long range object class is that it enables the mobilisation of spatial context, the details of which are explained in the following section. Ultimately, we aim to exploit colour band information and use spatial context to determine the real object class of long range targets.

6.5 Utilising Spatial Context

There is a wealth of evidence indicating the environment surrounding an object plays a very important part in human recognition systems [186, 187]. We aim to create a contextual framework to replicate this capability and provide accurate class information for long range targets, given their immediate scene context. We build on the previous work of Robertson et al. and extend their presented contextual framework [185]. The key components required for spatial context generation are:

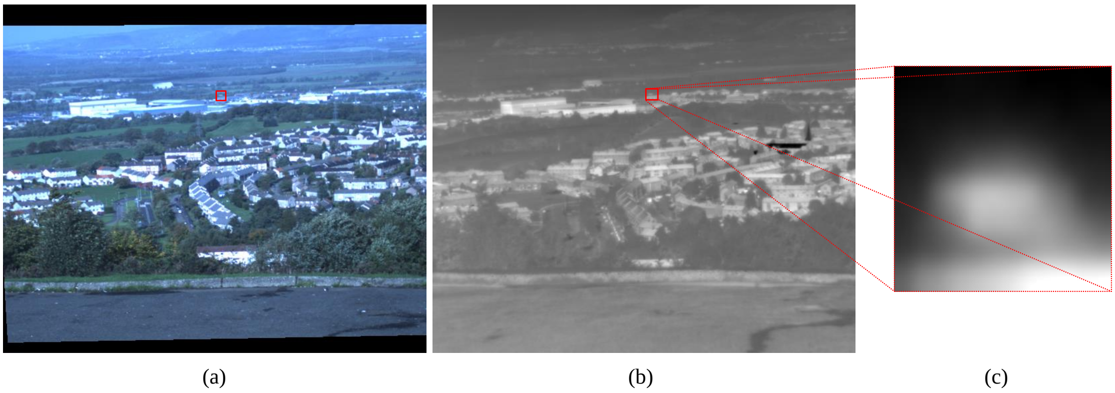


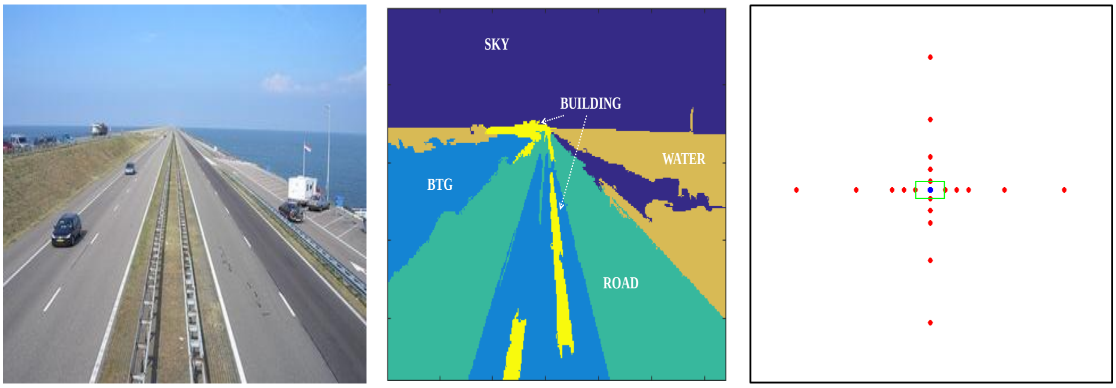
Figure 6.4: Long range sequence *The Braes*. Image (a) is a colour band image registered with central LWIR image (b). Both illustrate a candidate detection in the top, central portion of the image. The target is bounded by a red box. Image (c) is the target upsampled from the LWIR image information. Note that the detection algorithm only operates on the thermal image, but is shown on both colour and thermal for clarity [16].

a semantic segmentation algorithm, a spatial context sampling feature, prior scene knowledge for each object class and the probability for each object to exist given its surrounding context.

6.5.1 Semantic segmentation

A robust scene segmentation algorithm, capable of providing regions with semantic labels, provides a scene overview. We employ EGB segmentation to perform general region segmentation [2]. While this is a relatively old method it is still employed due to its simplicity at generating contiguous image regions, as discussed in Chapter 2. Semantic labeling is provided via a reimplementation of a recent method that also employs variations of EGB segmentation, but provides class labels via a trained SVM [4].

The SVM is trained using feature vectors composed of colour, texture and vertical position information from the segmented image regions. The colour features are computed as the mean and standard deviation for HSV colour planes, for each region. Texture features are generated by applying gabor filters at 1 scale and 8 orientations, to greyscale image regions. Assuming the (x, y) image coordinate system is used, the vertical position feature is simply the average value of *y-pixels* per region. We obtain image and ground truth data from the Stanford Background Dataset [188] to compute an array of these feature vectors to train our SVM, al-



(a) (b) (c)

Figure 6.5: A highway scene presented in image (a) is segmented and labeled with region class information, shown in image (b) with text overlay of the underlying semantic label. Image (c) is an illustration of the spatial context feature, for sampling the context surrounding an object using labeled images as shown in (b).

lowing segmented regions to be classified.

The image set contains 715 images, which we split into train/test using a $90 : 10$ ratio. The SVM can identify 5 distinct region classes: *sky, bush/tree/grass (BTG)* *, road, water* and *building.* This is a summary of the entire semantic segmentation method where output labeled regions provide spatial context for a scene. An example labeled segmentation is illustrated in Figure 6.5.

6.5.2 Context feature

In order to benefit from the semantic segmentation we require a sampling function to understand what regions surround candidate targets. We adopt the spatial context feature from the work of Robertson et al. to achieve this [185]. Given a detection with localisation information, we sample in four directions at five pixel locations. The directions are above, below, right and left of the detection centre. The pixels are sampled along each of these directions at increasing distances, $[1, 20, 40, 100, 200]$, starting from the edge of the detection bounding box.

It is clear that under this scheme there is a risk of sampling *off the image*, i.e. we are trying to sample the context at a pixel location that does not exist, especially if targets are observed at the edge of the image plane. To address this we pad the label image by replicating the boundary pixels for a fixed length in each direction, using the assumption that the regions continue indefinitely. For example, the sky region will continue above the image boundary. This circumvents any issues with context sampling and allows the computation of prior knowledge. The context sampling feature can be observed in subfigure (c) of Figure 6.5.

6.5.3 Prior knowledge

For each object detected in an image sequence we utilise the spatial context feature to sample surrounding context from labeled regions, which are output from the semantic segmentation algorithm. The region class at each sample point location can be observed and stored, building up a history of prior knowledge for each object class. This can be formulated in a probabilistic fashion, where we compute the prior knowledge of expected regions at sample locations per object class as $P_o(R_c|l_k)$. In this case P_o is the probability P per object class o, R_c is expected region R for region class c and l_k is sample l at location k.

The prior probabilities $P_o(R_c|l_k)$ for each object class, at each sample location, are learned normalised histograms from *observed* examples. To build this knowledge we obtain ≈ 20 random image examples from ImageNet [146] and DAGS [188] for each object class, which is 100 examples across the 5 real object classes. For each object class we semantically segment and sample the context around a target in the scene, using the results to compute the prior probability knowledge we require.

6.5.4 Probability of existence

We can use the prior information for expected regions in the following way. Given a colour and thermal band image sequence, we can obtain a candidate detection with corresponding localisation information in the thermal domain. Using the associated colour imagery, we can semantically segment and sample the context around the target. As mentioned earlier, each object class will have a learned probability for expected region at each sample location.

For example, the sample 100 pixels to the right of the target will have 5 expected region probabilities, for each object class. So if the sample 100 pixels to the right of the target returns a *sky* region, the learned probability for sky is taken. This value will be strongly linked to the object class, indicating where objects are most likely observed. Again we extend the work of Robertson's [185] contextual framework and formulate this as the probability of an object to exist given the surrounding spatial context is presented in Equation 6.1.

$$P(O|C) = \frac{1}{n} \sum P_o(R_c|l_k) \tag{6.1}$$

where $P(O|C)$ is the probability P for each object O to exist given the surrounding context C and n is number of sample points, which is 20 in our case. Put simply, the likelihood that an object would be found in a particular environment is computed as the average of all prior probabilities. Thus, for each target detection we will use Equation 6.1 to generate a 5 element array populated by the probability for each object to exist given the context. It is this array we will eventually use to affect the CNN output scores.

6.5.5 Temporal Aggregation

Although CNNs are generalisable and robust at classification tasks, incorrect classifications are still present. One possible method to address this is to informally track targets through a sequence and aggregate the output CNN scores over time, effectively squashing any erroneous probabilities. Considering the aim is to classify targets at long range, we can achieve temporal aggregation without implementing a tracking algorithm as far away targets move very little on the image plane.

This simple heuristic complements the long range target class which implicitly implies the observed object exists at a range that is unresolvable. Moreover, traditional tracking algorithms tend to perform data association via kinematics of targets, without considering any aspects of object recognition which our method incorporates. To implement this we formulate the problem using Bayes theorem to exploit the spatial and temporal relationship of targets in a video sequence. This can be summarised as trying to determine the probability of each object class Obj given previous classifications and detection locations through time T.

We create a circular buffer of detections with corresponding output CNN scores, where each output is a 1-dimensional vector or array of probability class scores $\mathbf{x}_{cnn}$. When a new detection and subsequent CNN score is acquired we can treat it is as an initial confidence of class values for that detection, which is *prior* knowledge $P(Obj)$. Using this we determine the posterior $P(Obj|T)$ by finding the closest spatial match in the circular buffer of detections, which serves as our likelihood function $P(T|Obj)$ to aggregate CNN class scores.

To compute the likelihood we search the previous detection locations in the buffer, finding the closest detection on the image plane in terms of Euclidean distance

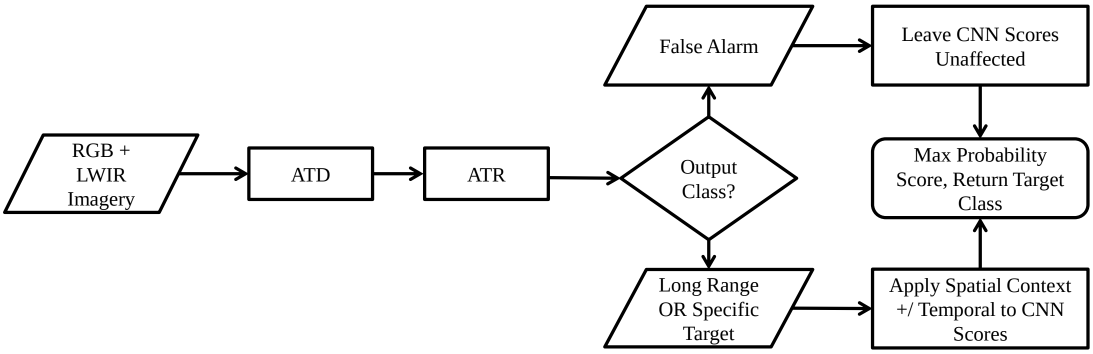

RGB + LWIR Imagery
ATD
ATR
Output Class?
False Alarm
Leave CNN Scores Unaffected
Max Probability Score, Return Target Class
Long Range OR Specific Target
Apply Spatial Context +/ Temporal to CNN Scores

Figure 6.6: General overview of the ATDR algorithm. Given multi-modal input data, candidate detections can be generated via an ATD process. These candidates are fed to the trained CNN, where the output class and score vector decides the next step. If the maximum class score is a false alarm, do nothing. If target is a long range class, remove FA and long range scores from CNN vector. Re-weight using spatial context. If real object class returned, re-weight CNN scores using spatial context.

det_{E_d}. If the nearest match is less than or equal to a defined distance threshold $Thresh_{E_d}$, i.e. spatially close, $P(T|Obj)$ becomes the corresponding CNN output vector containing all class scores. If the detection has no match below $Thresh_{E_d}$, current CNN output is unaffected. This condition is summarised in Equation 6.2:

$$P(T|Obj) = \begin{cases} \mathbf{x}_{cnn}, & \text{if } det_{E_d} \leq Thresh_{E_d} \\ 1, & \text{otherwise} \end{cases} \tag{6.2}$$

This gives us Bayes theorem as $P(Obj|T) \propto (T|Obj)P(Obj)$, which effectively describes how to update current target classes based on previous classifications, as well as spatial and temporal observations. The process then moves onto the next acquired detection and corresponding CNN scores, updating the circular buffer as required. By propagating through the detection sequence in this manner, the CNN scores are temporally aggregated and any spurious classifications diminished.

6.5.6 Final ATDR System

We can piece these components together for an effective end-to-end system and evaluate using real world, long range data. A graphical summary of how the key stages are linked is shown in Figure 6.6. For an input multi-modal data stream (RGB and LWIR) we generate candidate targets using Thales' proprietary ATD algorithm.

The corresponding image data for each target is upsampled, using the same process steps as employed for generating the CNN training data. The target crops can then be fed into the trained CNN and output classes reported.

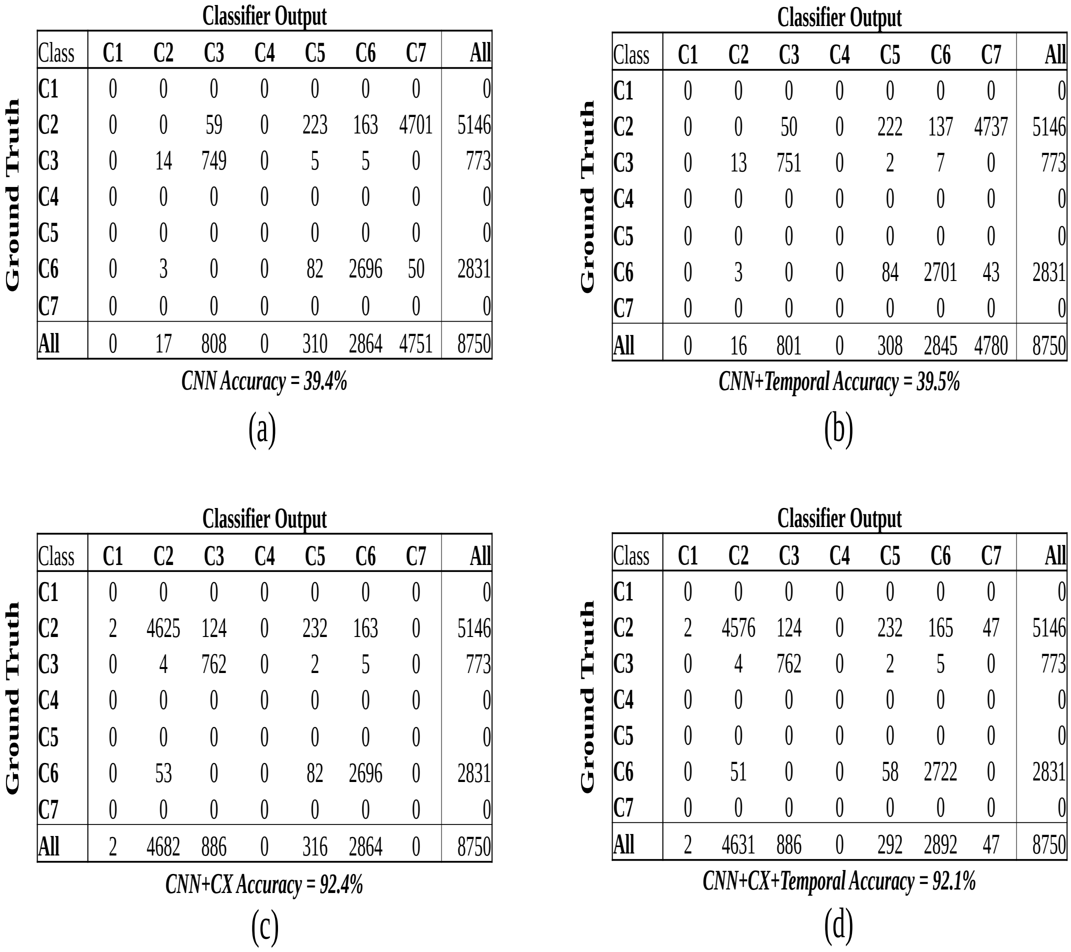
Upon obtaining the CNN output we have three possible options. If the top class returned is a false alarm, we simply leave the CNN score and record the label for this class. If it is a long range target, then we utilise localisation information from ATD and colour band imagery which is handed over to the contextual framework. This allows the probability for the target to be any of the object classes to be computed, returning a 5 element vector of probability scores PoC. Recall that our CNN is trained over 7 classes and consequently will output a 7 element score vector.

We have to remove the false alarm and long range target scores from the CNN vector, allowing CNN output and PoC probabilities to be merged. The resulting class scores are normalised so all entries sum to 1, the max of which provides the top class result used to evaluate the system. Lastly, if the maximum CNN score returns a real object class the system hands over target CNN scores to be affected by context in the same fashion.

6.6 Long Range Performance Evaluation

The final ATDR system is comprehensively evaluated using challenging, long range multi-modal data sequences collected in a rural location, illustrated in Figure 5.3. These scenes contained two main object classes, land vehicles and helicopter, as well as false alarms to classify. All detections generated via the ATD algorithm are human ground-truthed to provide target classes. Four combinations of the ATDR system are possible and we evaluate each of them, obtaining overall accuracy results and corresponding confusion matrices shown in Figure 6.7.

The simplest variation is *CNN*, where we apply only the trained CNN to targets output from ATD. *CNN+Temporal* applies the trained CNN and the temporal aggregation scheme outlined in Section 6.5.5. *CNN+CX* utilises spatial context to affect CNN scores and *CNN+CX+Temporal* applies every variation. The overall classifier accuracy for each combination is provided by Equation 6.3, where $Tr(cfm)$ is the trace of the confusion matrix and n_{cfm} is the total number of elements in the confusion matrix. We classify a total of 8750 long range target candidates for the final experiment



Classifier Output

Class	C1	C2	C3	C4	C5	C6	C7	All
C1	0	0	0	0	0	0	0	0
C2	0	0	59	0	223	163	4701	5146
C3	0	14	749	0	5	5	0	773
C4	0	0	0	0	0	0	0	0
C5	0	0	0	0	0	0	0	0
C6	0	3	0	0	82	2696	50	2831
C7	0	0	0	0	0	0	0	0
All	0	17	808	0	310	2864	4751	8750

CNN Accuracy = 39.4%

(a)

Classifier Output

Class	C1	C2	C3	C4	C5	C6	C7	All
C1	0	0	0	0	0	0	0	0
C2	0	0	50	0	222	137	4737	5146
C3	0	13	751	0	2	7	0	773
C4	0	0	0	0	0	0	0	0
C5	0	0	0	0	0	0	0	0
C6	0	3	0	0	84	2701	43	2831
C7	0	0	0	0	0	0	0	0
All	0	16	801	0	308	2845	4780	8750

CNN+Temporal Accuracy = 39.5%

(b)

Classifier Output

Class	C1	C2	C3	C4	C5	C6	C7	All
C1	0	0	0	0	0	0	0	0
C2	2	4625	124	0	232	163	0	5146
C3	0	4	762	0	2	5	0	773
C4	0	0	0	0	0	0	0	0
C5	0	0	0	0	0	0	0	0
C6	0	53	0	0	82	2696	0	2831
C7	0	0	0	0	0	0	0	0
All	2	4682	886	0	316	2864	0	8750

CNN+CX Accuracy = 92.4%

(c)

Classifier Output

Class	C1	C2	C3	C4	C5	C6	C7	All
C1	0	0	0	0	0	0	0	0
C2	2	4576	124	0	232	165	47	5146
C3	0	4	762	0	2	5	0	773
C4	0	0	0	0	0	0	0	0
C5	0	0	0	0	0	0	0	0
C6	0	51	0	0	58	2722	0	2831
C7	0	0	0	0	0	0	0	0
All	2	4631	886	0	292	2892	47	8750

CNN+CX+Temporal Accuracy = 92.1%

(d)

Figure 6.7: Confusion matrices and overall accuracy results are presented for long range classification experiments on the unseen test sets. **C1**= person, **C2** = land vehicle, **C3** = helicopter, **C4** = aeroplane, **C5** = UAV, **C6** = false alarm and **C7** = long range target class . Matrix (a) is simply the trained CNN applied to ATD output target candidates, which does not perform well. This result is almost identical when the temporal aggregation is introduced in matrix (b), with only a negligible gain on offer. However, context has an overwhelmingly positive affect on the ATDR results as shown in matrices (c) & (d).

.

$$Acc = \frac{Tr(cfm)}{n_{cfm}} \tag{6.3}$$

6.7 Discussion

The confusion matrices presented in Figure 6.3 highlight accurate classifier performance obtained via our CNN training scheme. Both overall accuracies exceed 95% , which suggests the additional seemingly uninformative $7th$ class for long range targets has no depreciable effect on recognition ability. The CNN structure of the classifier is highly symmetrical and deep, which also appears to be a suitable approach for our training dataset. Although we based our convnet architecture around Krizhevsky's successful ImageNet design, it is intriguing it translated so well given the network is designed for much larger scale learning.

ImageNet is $>$ 1 Million images over 1000 classes whereas our training set is approximately 10000 images distributed over 7 target classes. The ratio of 1000 images per class is roughly preserved and may be one reason why the network structure works well in both scenarios. In any case, the results obtained for LWIR target classification are significant and to our knowledge have never been shown before at this accuracy, or for the range of object classes we test over.

6.7.1 System Pros

Despite the effectiveness of the trained classifier, initial experiments presented in Section 6.4.2 highlight the ill-suited nature of CNN-only based classification of objects. Simply throwing a large scale machine learning approach at the long range problem is not viable for producing reliable target classes. However, these exploratory experiments identified the capability to differentiate false alarms from real targets, which is extremely useful and we capitalise on it by retraining our CNN with an additional class. This enables a Bayesian context framework to be employed.

By mobilising a very small amount of contextual information to affect CNN output scores, we gain reliable target class scores and demonstrate the benefit of such a system through extensive evaluation. The results of our final ATDR system, guided by context and temporal aggregation, are presented in Figure 6.7. We elect to keep all the confusion matrices in one figure for clarity as they show the whole picture. The experiments evaluate the end-to-end ATDR system over challenging long range sequences, generating 8750 target classes to examine. It is immediately clear from *CNN* and *CNN+Temporal* total accuracy results that the CNN is not able to discern the correct land vehicle classes. This will simply be due to the low signal quality at such long ranges.

Furthermore, the temporal aggregation function appears to have a negligible positive affect, only raising the overall classifier accuracy by a meager 0.1%. This may be explained by the fact that CNN output scores are very rarely *on the fence.* In other words they are very strong, consistently approaching $>$ 90% even when incorrect. Thus, even with temporal aggregation it makes it very hard to squash out the odd, erroneous classification. Out of thousands we only manage a handful, which is reflected in the negligible performance gain.

However, the performance improvement granted by the context framework is significant. Both algorithm variants including context achieve classification scores $\approx 92\%$, with the temporal function having a very slightly worse performance than just employing the CNN with context, although it is negligible again in terms of numbers. The improvement mostly comes from the correct switch from long range target to land vehicle, with slight gains in the helicopter class as well. The temporal function appears to be better at determining false alarms than the other variants, but it loses the accuracy gain elsewhere, mainly from preventing more long range target classes to switch to land vehicles. The entire experiment evaluation can be summarised by collating the CNN class scores into an $F_1 - Score$, which is given in Equation 6.4.

$$\begin{aligned} F_1 &= 2 \cdot \frac{precision \cdot recall}{precision + recall} \\ precision &= \frac{TP}{TP + FP} \\ recall &= \frac{TP}{TP + FN} \end{aligned} \tag{6.4}$$

This metric is a weighted average of *precision* and *recall* for binary classification, where TP is a true positive, FP is a false positive and TN is a true negative. Although ours is a multi-class problem we can still employ the $F_1 - Score$ per class and average over the results, giving a final quantitative summary of overall system performance. This summary can be visualised graphically and is shown in Figure 6.8. The illustration shows the average mean values for $F_1 - Scores$ and the corresponding percentage improvements, relative to the raw CNN output.

6.7.2 System Cons

A potential drawback of the system as a whole is the inability to affect incorrect false alarm cases, where the initial CNN output is misclassified as a false alarm, but they are relatively few so have little effect to overall accuracy. Furthermore, spatial context cannot resolve cases that are initially classed as objects but should be false alarms. This scenario is observed in Figure 6.7 (a)+(c), where the additional context has switched 50 long range class instances incorrectly to the land vehicle class. This is because there is no mechanism to enable target cases to become false

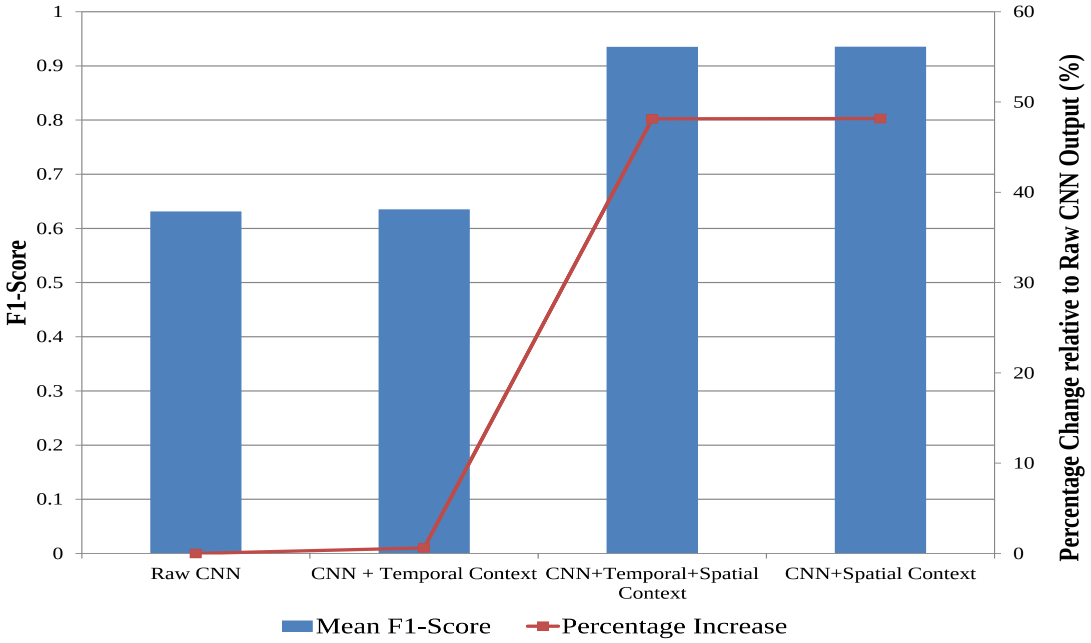


Figure 6.8: The multi-axis plot shows mean $F_1 - Scores$ for the different variants of classification algorithm in our final experiment. The $F_1 - Score$ is a useful summary statistic in machine learning as it provides an a weighted average of a classifiers precision and recall across classes. As we can see, there is a marked improvement gained from spatial context incorporation. This is highlighted by the red line showing the percentage increase in $F_1 - Score$ relative to the raw CNN output.

alarms via context, as it is not intuitive to create prior knowledge for a false alarm class where they could theoretically come from any region of an image. The system design enforces this rule.

Overall, by employing a very small amount of information, from semantic segmentation and object priors as context, we have successfully classified challenging long range targets in multi-modal surveillance data. Suffice to say, a little context goes a long way.

6.8 Conclusion

This chapter presents a complete ATDR system for enhancing target recognition capabilities in long range surveillance scenarios using challenging, real-world multi-modal data. This was achieved by initially adopting state-of-the-art machine learning methods to create a highly accurate LWIR target classifier via CNNs, demonstrating robust recognition across a range of objects in LWIR imagery. However, when this big data solution was shown to be inadequate for challenging long range scenarios, we mobilised a comparatively small amount of spatial context to infer accurate object classes. The approach is entirely data driven which will allow additional sensor information to be incorporated easily, either to classify a new target or improve the current system. We also demonstrated the ability to discriminate between real targets and false alarms to a high degree of accuracy. Building on this capability, the described ATDR system could potentially be deployed in a reconnaissance scenario and alleviate the burden on human operators via effective target reporting.

We illustrate how semantic region knowledge and the mobilisation of context, used in conjunction with the the presence of foreground objects, can be exploited to massively improve overall ATR performance for very challenging long range, multimodal surveillance data. Overall, the outlined approach should generalise very well to ATDR tasks in the security and defense domain, as well as outside this realm.

Overall, this chapter draws the experimental research portion of the thesis to a close. In the following final chapter we conclude the EngD by discussing the outcomes from the work and suggest future research directions.

Chapter 7
Conclusion

The grand aim of this thesis was to provide answers to the following interlinked questions:

1. *Can the presence of foreground objects influence scene perception and provide a route to persistent surveillance?*
2. *Does the mobilisation of scene specific context, gained via semantic segmentation, enhance target recognition performance using multimodal sensor systems?*

We have sufficiently addressed these questions and have determined the answer to both of them to be a resounding *yes*. We showed in Chapters 3 and 4 that the presence of foreground objects can be used to recover scene structure and classify thermal hotspots. However, the more useful and interesting discovery was discussed in Chapter 6, where we showed that the detection and discrimination of targets, which are foreground objects, from non-targets provides enhanced scene perception. It also enables the solution to the second question to be deployed.

The motivating factors driving the research are clearly outlined in Chapter 1, warrant asking such questions. The problem is considered from an academic and industrial viewpoint given the nature of the EngD itself, where we aspire to incorporate Thales manufactured sensors in proposed methods. In light of the engineering and industrial influence of the project, the overall solution had to be of relevance to the company and within the defence & security community. Moreover it must explicitly address the application we set out to improve, where we aim to provide improved situational awareness via enhancing scene segmentation and target recognition methodologies by mobilising contextual information.

Following the introductory motivational chapter we place the project in context by reviewing related literature across a wide range of fields within computer vision.

While the surveyed literature is broad, with only a few focus points in each section, we had to cover many different topics given the grand aim of the work undertaken. It encompasses object detection, segmentation and context methodologies, with an additional discussion on the growing relevance of machine learning within computer vision. We describe the key supervised learning tools used in our work and present recent SOA examples for object detection and segmentation that leverage this powerful technique. We conclude Chapter 2 by initiating a discussion on the tightly coupled relationship between foreground and background regions.

Chapter 3 begins our experimental research work by examining this relationship. We use a typical surveillance scenario and corresponding colour dataset. The imagery contains crowded urban scenes from a static camera viewpoint. We develop a method that exploits emergent foreground context, in the form of people tracks through the scene, to recover underlying scene structure that is otherwise unobtainable. While the presented solution is based on the assumption that the target scenes are always crowded with foreground objects, it was a useful exercise to prove the core idea underpinning the method. We effectively wanted to show that the real world and interactions within are quite structured, behaving in a certain way. The presence of foreground object inherently tells us something important about the scene itself and the surrounding environment without directly measuring it. This idea is carried forward to Chapter 4 where we introduce thermal sensor data to begin examining the 24-hour surveillance problem.

In Chapter 4 we discuss two region segmentation methods for characterising scene structure, using colour and thermal band imagery. The work develops a technique to identify and classify thermal signatures in the multi-modal footage, providing the means to offer true *round the clock* surveillance capabilities. The first of these segmentation methods was an adaption of the algorithm presented in chapter 3, where the thermal channel data was incorporated into the existing region merging framework The alternative method was a simple, novel design tracking variance through superpixels. This method enabled the presence of foreground objects to be accurately observed. A knowledge transfer scheme based on Bayesian mathematics was enforced and it enabled thermal signatures to be classified without a bespoke trained classifier. We showed this for an open source benchmark dataset as well as using our own real world multi-modal data. However, despite the reasonable performance of the system it is obviously a cumbersome solution. We address this by investigating the rapidly ascending world of machine learning and CNNs towards LWIR object classification.

Chapters 5 and 6 are closely linked and help to answer our stated goals. We began by collecting vast amounts of LWIR imagery using the high quality TI manufactured by Thales, the Catherine MP. This bounty of data was transformed into a fully labelled dataset after many man hours spent cropping out defined targets, labelling them and applying a preprocessing step. Once we obtained this valuable dataset we could finally apply CNNs and efficient optimisation schemes to create a high performance target classifier. The research geared towards machine learning and network behaviour is explored in Chapter 5. We then take this trained CNN and embed it within an end-to-end model, applying it to challenging real world surveillance data relevant to Thales. The application of the CNN and spatial context towards enhancing ATR performance is fully explored in Chapter 6.

Finally, in Chapter 6, we arrive at the major problem embodied within this thesis. The multi-modal data contains real world surveillance footage containing long range targets in rural environments, a typical scenario in battlespace environments where improved situational awareness is vital. We develop an effective solution by adapting the CNN from Chapter 5 to include a false alarm class, which is incredibly accurate at target discrimination. We also encode object priors and a region segmentation method to guide CNN class output. By combining all these elements we demonstrated our ATDR system was very effective at discerning target class for long range signatures, that would certainly be impossible to otherwise. In doing so we can confidently state that *yes - the presence of foreground objects does influence scene perception, providing a route to persistent surveillance* and *yes - mobilising scene specific context does enhance target recognition performance utilising multi-modal sensor systems.*

7.1 Research Contributions

We can summarise the novel contributions of our work in the following ways. A scene segmentation algorithm influenced by the presence of pedestrians, as well as other foreground objects, is proposed. It is applicable to colour band surveillance imagery. How the algorithm affects the underlying background structure of a scene is ascertained through extensive evaluation. Furthermore, the segmentation process is adapted to incorporate thermal band data providing a route to 24-hour surveillance. This algorithm and a similar superpixel based effort are employed to effectively gain target classification from thermal imagery via domain knowledge transfer. The work was presented at the IEEE conference, International Conference on Image Processing in 2015 [23].

We create an extensive labelled object dataset for LWIR imagery, collected using the Thales Catherine MP sensor. This enabled a high performance convolutional neural network to be developed for target classification purposes, where we are one of the first to achieve this feat. The robust algorithm is applied to a variety of scenarios that Thales operate in, with the most challenging being long range detection and recognition. We develop a sufficient model for enhancing long range LWIR target classification using scene specific context, for the representative problem data collected by a Thales sensor. Performance is verified through extensive evaluation. Some of this work has been presented at the SPIE conference on Security & Defense in 2016 [16]. The remainder of the work is due to be presented at SSPD London in 2017.

7.2 Commercial Contributions and Impact

The EngD must not only produce novel research in academic terms, it has to have an impact within industry as well. Towards this we produced an outcome with significant impact during the project, implementing a real-time CNN demo for target classification as described in Chapter 5, specifically Figure 5.15. The demo was showcased at a high profile internal research event where the stall won *best innovation* and is directly relevant to many Thales products.

Furthermore, this project has changed the direction of research strategy and development within Thales Land and Air Systems based on the results described throughout the thesis. This is a concise summation of what an EngD project should achieve in terms of commercial impact.

7.3 Future Work

As always there are many areas of interest that could be investigated further. However, we shall only cover the most interesting of these and address what we see as a possible gap.

The most obvious next step would be to advance the work presented in Chapter 6 to include anomaly detection experiments. The algorithm is effective for long range target classification as it exploits the defined relationships between spatial context and object prior information. In other words it uses the evidence of what regions are objects likely to appear in to guide CNN output. Without much more alteration to the model the question can effectively be asked in reverse - *can we exploit the knowledge of where objects are likely to appear, to identify when targets are out of context?*. All that would be required is a suitable dataset with anomalous behaviour to observe, then strong CNN classification output can be flagged if the object appears in an unlikely area etc.

Lastly, a natural advancement of the work would be to incorporate colour band data into our thermal CNN classifier. The general idea would require a registered set of multi-modal imagery, with objects appearing in both sensors concurrently. This would enable a new labelled dataset to be created showing how the same object appears in multiple sensor spaces. If network architecture could be successfully navigated to allow optimisation, then we would gain a very robust target classifier for persistent surveillance. It would also simply algorithm design as both sensor streams could be routed through the single trained CNN, which can run day or night, providing accurate target recognition capability.

References

[1] A. Rankin, A. Huertas, L. Matthies, M. Bajracharya, C. Assad, S. Brennan, P. Bellutta, and G. W. Sherwin, "Unmanned ground vehicle perception using thermal infrared cameras," in *SPIE Defense, Security, and Sensing.* International Society for Optics and Photonics, 2011, p. 804503. xii, xvi, 18, 49, 74

[2] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004. xii, 25, 28, 29, 65, 133

[3] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012. xii, 29, 30, 31, 59

[4] M. A. Pieck, F. van der Sommen, S. Zinger, and P. H. de With, "Real-time semantic context labeling for image understanding," in *Image Processing (ICIP), 2015 IEEE International Conference on.* IEEE, 2015, pp. 3180–3184. xiii, 25, 34, 35, 133

[5] J. Kumagai, "A robotic sentry for korea's demilitarized zone," *IEEE Spectrum*, vol. 44, no. 3, pp. 16–17, 2007. xiii, 37

[6] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004. xiii, 38, 40, 41

[7] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba, "Inverting and visualising features for object detection," *arXiv preprint arXiv:1212.2278*, 2012. xiii, 42

[8] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool, "Pedestrian detection at 100 frames per second," in *Computer Vision and Pattern Recognition*

(CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 2903–2910. xiv, 38, 43, 44

[9] A. Torralba, "Contextual priming for object detection," *International journal of computer vision*, vol. 53, no. 2, pp. 169–191, 2003. xiv, 52, 53

[10] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001. xiv, 52, 53

[11] "CAVIAR: Context aware vision using image-based active recognition edinburgh university informatics department," [Online] Available: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/. xiv, xv, 57, 67

[12] B. Benfold and I. Reid, "Stable multi-target tracking in real - time surveillance video," in *Computer Vision and Pattern Recogntion*, 2011, pp. 3457–3464. xiv, 19, 57, 58

[13] "Using thermal cameras to secure the homeland," 2010. [Online]. Available: https://www.photonics.com/Article.aspx?AID=40915 xvi, 74, 75

[14] J. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *IEEE OTCBVS WS Series Bench; Computer Vision and Image Understanding*, vol. 106, pp. 162 – 182, 2007. xvi, 85, 101

[15] A. Berg, J. Ahlberg, and M. Felsberg, "A thermal object tracking benchmark," in *Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on.* IEEE, 2015, pp. 1–6. xvii, 19, 48, 98

[16] I. Rodger, B. Connor, and N. M. Robertson, "Classifying objects in lwir imagery via cnns," in *SPIE Security+ Defence.* International Society for Optics and Photonics, 2016, p. 99870H. xvii, xix, xx, 9, 92, 99, 104, 126, 130, 131, 132, 133, 147

[17] J. Byrnes, *Unexploded ordnance detection and mitigation.* Springer Science & Business Media, 2008. xxii, 18

[18] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on.* IEEE, 2009, pp. 1271–1278. xxii, 50, 51

[19] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 37, no. 1, pp. 32–64, 1995. 1, 5, 93

[20] S. Crawford, R. Craig, A. Haining, J. Parsons, E. Costard, P. Bois, F.-H. Gauthier, and O. Cocle, "Thales long-wave advanced ir qwip cameras," in *Defense and Security Symposium.* International Society for Optics and Photonics, 2006, p. 62060H. 6, 84, 93, 102

[21] A. Pimenta, D. Carneiro, P. Novais, and J. Neves, "Analysis of human performance as a measure of mental fatigue," in *International Conference on Hybrid Artificial Intelligence Systems.* Springer, 2014, pp. 389–401. 7

[22] S. Li, W. Chen, Y. Fu, C. Wang, Y. Tian, and Z. Tian, "Investigating the effects of experience on human performance in an object-tracking task: a case study of manual rendezvous and docking," *Behaviour & Information Technology*, vol. 35, no. 6, pp. 427–441, 2016. 7

[23] I. Rodger, B. Connor, and N. M. Robertson, "Recovering background regions in videos of cluttered urban scenes," in *Image Processing (ICIP), 2015 IEEE International Conference on.* IEEE, 2015, pp. 4097–4101. 9, 56, 76, 147

[24] I. Rodger, B. Connor, R. Abbott, and N. M. Robertson, "Enhancing long range atr using spatial context [to be presented]," in *Sensor Signal Processing for Defence (SSPD).* IEEE, 2017. 9

[25] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis and Machine Vision.* Chapman & Hall, 1993. 15

[26] A. M. Pinto, L. F. Rocha, and A. P. Moreira, "Object recognition using laser range finder and machine learning techniques," *Robotics and Computer-Integrated Manufacturing*, pp. 12–22, 2013. 15

[27] A. Hattori, A. Hosaka, M. Taniguchi, and E. Nakano, "Driving control system for an autonomous vehicle using multiple observed point information." in *Proc. Intelligent Vehicles Symposium*, 1992, pp. 207–212. 15

[28] B. B. et al, "Introduction to the special issue on learning in computer vision and pattern recognition," *IEEE Transactions on Systems, Man and Cybernetics*, June 2005. 15

[29] H. D. Young, R. A. Freedman, and L. Ford., *University Physics.* Addison-Wesley, 2007. 16

[30] R. M. White, "A sensor classification scheme," *IEEE Transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 34, no. 2, pp. 124–126, 1987. 16

[31] J. Sugden, “Watched by the games: Surveillance and security at the olympics.” *Int. Review for the Sociology of Sport.*, vol. 47, no. 3, pp. 414–429, 2012. 16

[32] J. M. Kasson and W. Pluffe, “An analysis of selected computer interchange color spaces,” in *ACM Transactions on Graphics*, vol. 11, no. 2, 1992, pp. 373–405. 16

[33] K. Asakawa and H. Sugiura, “High-precision colour transformation system.” in *IEEE Transactions on Consumer Electronics*, vol. 41, no. 2, 1995, pp. 373–405. 16

[34] L. Li and F. Y. Wang, “Intelligent vehicle vision systems,” *Springer - Advanced Motion Control and Sensing for Intelligent Vehicles*, pp. 323–399, 2007. 17

[35] C. M. Johnston, N. A. Mould, and J. P. Havlicek., “Multichannel dual domain infrared target tracking for highly evolutionary target signatures,” in *IEEE Int. Conf. on Image Processing*, 2009. 18

[36] “Catherine mp mw technical specification,” Tech. Rep., http://www.thalesgroup.com/Countries/United_Kingdom/Sensors_Microsite/Products/Sensors_Tech_Spec_pg_Catherine_MP_MW/. 18

[37] J. W. Davis and V. Sharma, “Fusion-based background-subtraction using contour saliency,” in *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on.* IEEE, 2005, pp. 11–11. 19

[38] ——, “Background-subtraction using contour-based fusion of thermal and visible imagery,” *Computer Vision and Image Understanding*, vol. 106, no. 2, pp. 162–182, 2007. 19, 84

[39] J. R. Searle, “The chinese room revisited,” *Behavioral and brain sciences*, vol. 5, no. 02, pp. 345–348, 1982. 19

[40] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, “Mastering the game of go with deep neural networks and tree search,” *Nature*, vol. 529, no. 7587, pp. 484–489, 2016. 19

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105. 19, 44, 45, 108, 114

[42] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," pp. 3–24, 2007. 20

[43] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE transactions on pattern analysis and machine intelligence*, vol. 19, no. 2, pp. 153–158, 1997. 20, 95

[44] B. J. Schachter, "Target classification strategies," in *SPIE Defense+ Security*. International Society for Optics and Photonics, 2015, p. 947602. 21, 95

[45] K. P. Murphy, "Machine learning: a probabilistic perspective," p. 2, 2012. 21, 96

[46] B. Cheng and D. M. Titterington, "Neural networks: A review from a statistical perspective," *Statistical science*, pp. 2–30, 1994. 23

[47] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 1, pp. 4–37, 2000. 23

[48] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015. 23, 94

[49] R.M.Haralick and L. G. Shapiro, *Computer and Robot Vision, Vols I and II.* Addison-Wesley, 1992. 24

[50] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona, "What do we perceive in a glance of a real-world scene?" *Journal of vision*, vol. 7, no. 1, pp. 10–10, 2007. 24

[51] G. A. Hance, S. E. Umbaugh, R. H. Moss, and W. V. Stoecker, "Unsupervised color image segmentation: with application to skin tumor borders," *IEEE Engineering in Medicine and Biology Magazine*, vol. 15, no. 1, pp. 104–111, 1996. 24

[52] M.-N. Wu, C.-C. Lin, and C.-C. Chang, "Brain tumor detection using color-based k-means clustering segmentation," in *Intelligent Information Hiding and Multimedia Signal Processing, 2007. IIHMSP 2007. Third International Conference on*, vol. 2. IEEE, 2007, pp. 245–250. 24

[53] T. Kaur, B. S. Saini, and S. Gupta, "Optimized multi threshold brain tumor image segmentation using two dimensional minimum cross entropy based on co-occurrence matrix," in *Medical Imaging in Clinical Applications.* Springer, 2016, pp. 461–486. 24

[54] H. Zhou, H. Kong, L. Wei, D. Creighton, and S. Nahavandi, “On detecting road regions in a single uav image,” *IEEE Transactions on Intelligent Transportation Systems*, 2016. 24

[55] M. Li, A. Stein, W. Bijker, and Q. Zhan, “Region-based urban road extraction from vhr satellite images using binary partition tree,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 44, pp. 217–225, 2016. 24

[56] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, “Contour detection and hierarchical image segmentation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 898–916, 2011. 25

[57] B. Fulkerson, A. Vedaldi, and S. Soatto, “Class segmentation and object localization with superpixel neighborhoods,” in *Computer Vision, 2009 IEEE 12th International Conference on.* IEEE, 2009, pp. 670–677. 29

[58] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa, “Entropy rate superpixel segmentation,” in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on.* IEEE, 2011, pp. 2097–2104. 29

[59] J. Shen, Y. Du, W. Wang, and X. Li, “Lazy random walks for superpixel segmentation,” *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1451–1462, 2014. 29

[60] X. Ren and J. Malik, “Learning a classification model for segmentation.” in *International Conference for Computer Vision*, vol. 1, 2003, pp. 10–17. 29, 32, 60

[61] A. Levinshtein, A. Stere, K. Kutulakos, D. Fleet, S. Dickinson, and K. Siddiqi, “Turbopixels: Fast superpixels using geometric flows,” *PAMI*, vol. 31, no. 12, pp. 2290–2297, 2009. 29

[62] C. Connolly and T. Fleiss, “A study of efficiency and accuracy in the transformation from rgb to cielab color space,” *IEEE Transactions on Image Processing*, vol. 6, no. 7, pp. 1046–1048, 1997. 30

[63] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman., “The pascal visual object classes challenge results,” *Int. Journal of Computer Vision.*, vol. 88, no. 2, pp. 303–338, 2010. 32, 87

[64] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, “Conditional random fields as recurrent neural

networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1529–1537. 32, 33

[65] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *arXiv preprint arXiv:1606.00915*, 2016. 33

[66] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 33

[67] F. Liu, G. Lin, and C. Shen, "Crf learning with cnn features for image segmentation," *Pattern Recognition*, vol. 48, no. 10, pp. 2983–2992, 2015. 33

[68] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528. 33

[69] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1713–1721. 33

[70] J. Letham, N. M. Robertson, and B. Connor, "Contextual smoothing of image segmentation," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on.* IEEE, 2010, pp. 7–12. 34

[71] P. Dollár, C. wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evluation of the state of the art," *IEEE Transactions on PAMI*, 2012. 37

[72] P. Dollár, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in *British Machine Vision Conference*, 2010. 38, 44

[73] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.* IEEE, 2008, pp. 1–8. 38

[74] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893. 38, 40, 97

[75] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 11, pp. 1475–1490, 2004. 38

[76] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *International journal of computer vision*, vol. 38, no. 1, pp. 15–33, 2000. 38

[77] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010. 38, 40

[78] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 3. IEEE, 2006, pp. 850–855. 40

[79] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587. 45

[80] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448. 46

[81] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 91–99. 46

[82] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 46

[83] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," *arXiv preprint arXiv:1612.08242*, 2016. 46

[84] H. Abramson and S. Avidan, "Tracking through scattered occlusion," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on.* IEEE, 2011, pp. 1–8. 47

[85] M. I. Ribeiro, "Kalman and exteneded kalman filters: concept, derivation and properties." Institute for Systems and Robotics, Tech. Rep., 2004. 47

[86] K. Okuma, A. Taleghani, N. d. Freitas, J. J. Little, and D. G. Lowe, "A boosted particle filter: Multitarget detection and tracking," *Computer Vision-ECCV 2004*, pp. 28–39, 2004. 47

[87] B. Connor, I. Carrie, R. Craig, and J. Parsons, "Discriminative imaging using a lwir polarimeter," *Electro-Optical and Infrared Systems: Technology and Applications V*, 2008. 48

[88] Z. Yin and R. Collins, "Moving object localization in thermal imagery by forward-backward mhi," in *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on.* IEEE, 2006, pp. 133–133. 48

[89] W. Wang, J. Zhang, and C. Shen, "Improved human detection and classification in thermal images," in *Proc. IEEE International Conference on Image Processing*, 2010. 48

[90] J. Davis and M. Keck, "A two-stage template approach to person detection in thermal imagery," in *IEEE Workshop: Application of Computer Vision*, 2005, pp. 364–369. 48

[91] A. Berg, M. Felsberg, G. Häger, and J. Ahlberg, "An overview of the thermal infrared visual object tracking vot-tir2015 challenge," in *Swedish Symposium on Image Analysis*, 2016. 48

[92] A. Berg, J. Ahlberg, and M. Felsberg, "Channel coded distribution field tracking for thermal infrared imagery," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016. 48

[93] P. Harding and N. M. Robertson, "Visual saliency from image features with application to compression," *Cognitive Computing*, vol. 5, no. 1, pp. 76–98, 2013. 49

[94] B. C. Hansen and E. A. Essock, "A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes," *Journal of vision*, vol. 4, no. 12, pp. 5–5, 2004. 50

[95] T. M. Strat, "Employing contextual information in computer vision," *DARPA93*, pp. 217–229, 1993. 50, 128

[96] I. Biederman, "Recognition-by-components: a theory of human image understanding." *Psychological review*, vol. 94, no. 2, pp. 115–148, 1987. 51

[97] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz, "Scene perception: Detecting and judging objects undergoing relational violations," *Cognitive psychology*, vol. 14, no. 2, pp. 143–177, 1982. 51

[98] I. Biederman, "Human image understanding: Recent research and a theory," *Computer vision, graphics, and image processing*, vol. 32, no. 1, pp. 29–73, 1985. 51

[99] A. Friedman, "Framing pictures: the role of knowledge in automatized encoding and memory for gist," *Journal for Experimental Psychology: General*, vol. 108, pp. 316–355, 1979. 52

[100] A. Borji, S. Frintrop, D. N. Sihite, and L. Itti, "Adaptive object tracking by learning background context," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on.* IEEE, 2012, pp. 23–30. 53

[101] E. Maggio and A. Cavallaro, "Learning scene context for multiple object tracking," in *IEEE Transactions on Image Processing*, vol. 18, no. 8, 2009, pp. 1873–1884. 54

[102] R. S. Medeiros, J. Scharcanski, and A. Wong, "Natural scene segmentation based on a stochastic texture region merging approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 1464–1467. 58

[103] X. Wang, C. Zhu, C. E. Bichot, and S. Masnou, "Graph-based Image Segmentation Using Weighted Color Patch," in *IEEE International Conference on Image Processing (ICIP)*, 2013, pp. 4064–4068. 58

[104] J. Letham, N. Robertson, and B. Connor, "Contextual smoothing of image segmentation," in *Computer Vision and Pattern Recognition Workshops, 2010 IEEE Computer Society Conference on.* IEEE, 2010, pp. 7–12. 58, 59, 64

[105] Y. Tian, Y. Wang, Z. Hu, and T. Huang, "Selective eigenbackground for background modeling and subtraction in crowded scenes," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 23, no. 11, pp. 1849–1864, 2013. 58

[106] S. Gould, T. Gao, and D. Koller, "Region-based segmentation and object detection," in *Advances in neural information processing systems*, 2009, pp. 655–663. 58, 59

[107] C. Wojek, "A dynamic conditional random field model for joint labeling of object and scene classes." in *European Conference on Computer Vision.* Springer, 2008, pp. 733–747. 58

[108] C. Gu, J. J. Lim, P. Arbelaez, and J. Malik, “Recognition using regions,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on.* IEEE, 2009, pp. 1030–1037. 58

[109] B. Liu, S. Gould, and X. He, “Multi-class semantic video segmentation with exemplar-based object reasoning,” in *Winter Conference on Applications of Computer Vision*, 2015. 59

[110] D. Sun, J. Wulff, B. Sudderth, H. Pfister, and M. J. Black, “A fully-connected layered model of foreground and background flow,” in *Computer Vision and Pattern Recognition, CVPR 2013, IEEE Conference on.* IEEE, 2013, pp. 2451–2458. 59

[111] F. Navarro, M. Escudero-Viñolo, and J. Bescos, “Sp-sift: Enhancing sift discrimination via super-pixel-based foreground-background segregation,” *Electronic Letters*, vol. 50, no. 4, pp. 272–274, 2014. 59

[112] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi, “Turbopixels: Fast superpixels using geometric flows,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 31, no. 12, pp. 2290–2297, 2009. 60

[113] F. J. Aherne, N. A. Thacker, and P. I. Rockett, “The bhattacharyya metric as an absolute similarity measure for frequency coded data,” *Kybernetika*, vol. 34, no. 4, pp. 363–368, 1998. 60

[114] D. Comaniciu and P. Meer, “Mean shift: A robust approach toward feature space analysis,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002. 60

[115] R. H. Baxter, M. J. Leach, S. S. Mukherjee, and N. M. Robertson, “An adaptive motion model for person tracking with instantaneous head-pose features,” *IEEE Signal Processing Letters*, vol. 22, no. 5, pp. 578–582, 2015. 62

[116] N. Anjum and A. Cavallaro, “Multifeature object trajectory clustering for video analysis,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1555–1564, 2008. 62, 65

[117] A. Fitzgibbon, M. Pilu, and R. B. Fisher, “Direct least square fitting of ellipses,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 21, no. 5, pp. 476–480, 1999. 62

[118] V. Epanechnikov, “Non-parametric estimation of a multivariate probability density,” *Theory of Probability & Its Applications*, vol. 14, no. 1, pp. 153–158, 1969. 63

[119] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000. 65

[120] M. Meilă, “Comparing clusterings by the variation of information,” in *Learning theory and kernel machines.* Springer, 2003, pp. 173–187. 65

[121] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, “From contours to regions: An empirical evaluation,” in *Computer Vision and Pattern Recognition, 2009.* IEEE, 2009, pp. 2294–2301. 66

[122] S. L. Dockstader, M. J. Berg, and A. M. Tekalp, “Stochastic kinematic modeling and feature extraction for gait analysis,” *IEEE Transactions on Image Processing*, vol. 12, no. 8, pp. 962–976, 2003. 73

[123] T. E. Boult, X. Gao, R. Micheals, and M. Eckmann, “Omni-directional visual surveillance,” *Image and Vision Computing*, vol. 22, no. 7, pp. 515–534, 2004. 73

[124] R. Feris, B. Siddiquie, Y. Zhai, J. Petterson, L. Brown, and S. Pankanti, “Attribute-based vehicle search in crowded surveillance videos,” in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval.* ACM, 2011, p. 18. 73

[125] F. Porikli, F. Brémond, S. L. Dockstader, J. Ferryman, A. Hoogs, B. C. Lovell, S. Pankanti, B. Rinner, P. Tu, and P. L. Venetianer, “Video surveillance: past, present, and now the future [dsp forum],” *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 190–198, 2013. 73

[126] G. D. Hager and P. N. Belhumeur, “Efficient region tracking with parametric models of geometry and illumination,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 10, pp. 1025–1039, 1998. 73

[127] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, “Robust real-time unusual event detection using multiple fixed-location monitors,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 3, pp. 555–560, 2008. 73

[128] (2017) FLIR commercial vision systems. [Online]. Available: http://www.flir.co.uk/cs/display/?id=51839 75

[129] P. Dollár, R. Appel, S. Belongie, and P. Perona, “Fast feature pyramids for object detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014. 77

[130] E. Ohn-Bar and M. M. Trivedi, “Learning to detect vehicles by clustering appearance patterns,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2511–2521, 2015. 77

[131] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions,” *Image and vision computing*, vol. 22, no. 10, pp. 761–767, 2004. 82, 96

[132] M. Teutsch, T. Muller, M. Huber, and J. Beyerer, “Low resolution person detection with a moving thermal infrared camera by hot spot classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 209–216. 82

[133] A. Papoulis and S. U. Pillai, *Probability, random variables, and stochastic processes.* Tata McGraw-Hill Education, 2002. 83

[134] P. Meer, “Robust techniques for computer vision,” *Emerging topics in computer vision*, pp. 107–190, 2004. 93

[135] B. Bhanu, J. Peng, T. Huang, and B. Draper, “Introduction to the special issue on learning in computer vision and pattern recognition,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 35, no. 3, pp. 391–396, 2005. 93

[136] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. 93

[137] L. Becker, “Influence of IR sensor technology on the military and civil defense,” vol. 6127, 2006, p. 61270S. 93

[138] T. P. Breckon, J. W. Han, and J. Richardson, “Consistency in multi-modal automated target detection using temporally filtered reporting,” in *SPIE Security+ Defence.* International Society for Optics and Photonics, 2012, p. 85420L. 93

[139] A. Leykin and R. Hammoud, “Robust multi-pedestrian tracking in thermal-visible surveillance videos,” in *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW’06).* IEEE, 2006, pp. 136–136. 93

[140] A. Clapés, M. Reyes, and S. Escalera, "Multi-modal user identification and object recognition surveillance system," *Pattern Recognition Letters*, vol. 34, no. 7, pp. 799–808, 2013. 93

[141] J. Yin, L. Liu, H. Li, and Q. Liu, "The infrared moving object detection and security detection related algorithms based on w4 and frame difference," *Infrared Physics & Technology*, vol. 77, pp. 302–315, 2016. 93

[142] B. P. Bailey and J. A. Konstan, "On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state," *Computers in human behavior*, vol. 22, no. 4, pp. 685–708, 2006. 93

[143] R. Hockey, *The psychology of fatigue: work, effort and control.* Cambridge University Press, 2013. 93

[144] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012. 93, 110

[145] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989. 93

[146] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015. 94, 98, 135

[147] 2016. [Online]. Available: http://karpathy.github.io/2015/10/25/selfie/ 94

[148] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–511. 96

[149] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1. IEEE, 2002, pp. I–900. 96

[150] T. P. Breckon, A. Gaszczak, J. Han, M. L. Eichner, and S. E. Barnes, "Multimodal target detection for autonomous wide area search and surveillance," in *SPIE Security+ Defence.* International Society for Optics and Photonics, 2013, p. 889913. 96, 127, 128

[151] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999. 96

[152] M. Teutsch, T. Muller, M. Huber, and J. Beyerer, "Low resolution person detection with a moving thermal infrared camera by hot spot classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 209–216. 97, 99, 130

[153] A. S. Arya, D. T. Anderson, C. L. Bethel, and D. Carruth, "Multi-kernel aggregation of local and global features in long-wave infrared for detection of swat teams in challenging environments," vol. 8744, 2013, p. 87440O. 97

[154] H. K. Ekenel and R. Stiefelhagen, "Local appearance based face recognition using discrete cosine transform," in *13th European Signal Processing Conference (EUSIPCO 2005), Antalya, Turkey*, 2005. 97

[155] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," pp. 91.1–91.11, 2009. 97

[156] A. Prinzie and D. Van den Poel, "Random multiclass classification: generalizing random forests to random mnl and random nb," in *International Conference on Database and Expert Systems Applications*. Springer, 2007, pp. 349–358. 97

[157] K. Stone, J. M. Keller, D. T. Anderson, and D. B. Barclay, "An automatic detection system for buried explosive hazards in fl-lwir and fl-gpr data," vol. 8357, 2012, p. 83571E. 97

[158] S. R. Price, D. T. Anderson, R. H. Luke, K. Stone, and J. M. Keller, "Automatic detection system for buried explosive hazards in fl-lwir based on soft feature extraction using a bank of gabor energy filters," vol. 8709, 2013, p. 87091B. 97

[159] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996. 97

[160] B. Li, R. Chellappa, Q. Zheng, S. Der, N. Nasrabadi, L. Chan, and L. Wang, "Experimental evaluation of flir atr approaches: A comparative study," *Computer Vision and image understanding*, vol. 84, no. 1, pp. 5–24, 2001. 99

[161] K. Stone and J. Keller, "Convolutional neural network approach for buried target recognition in fl-lwir imagery," in *SPIE Defense+ Security*. International Society for Optics and Photonics, 2014, p. 907219. 99

[162] M. Chevalier, N. Thome, M. Cord, J. Fournier, G. Henaff, and E. Dusch, "Low resolution convolutional neural network for automatic target recognition," in *7th International Symposium on Optronics in Defence and Security*, 2016. 99

[163] E. J. Lee, B. C. Ko, and J.-Y. Nam, "Recognizing pedestrians unsafe behaviors in far-infrared imagery at night," *Infrared Physics & Technology*, vol. 76, pp. 261–270, 2016. 99

[164] C. Herrmann, T. Müller, D. Willersinn, and J. Beyerer, "Real-time person detection in low-resolution thermal infrared imagery with mser and cnns," in *SPIE Security+ Defence*. International Society for Optics and Photonics, 2016, p. 99870I. 99, 101

[165] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, University of Cambridge, 2016. 103

[166] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009. 105

[167] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4, pp. 411–430, 2000. 107

[168] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814. 108

[169] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks." in *Aistats*, vol. 9, 2010, pp. 249–256. 109

[170] J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neurocomputing*. Springer, 1990, pp. 227–236. 110

[171] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011. 110

[172] X. Li and J. Bilmes, "Regularized adaptation of discriminative classifiers," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1. IEEE, 2006, pp. I–I. 111

[173] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. 114

[174] S. S. Mukherjee and N. M. Robertson, “Deep head pose: Gaze-direction estimation in multimodal video,” *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2094–2107, 2015. 116

[175] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” *arXiv preprint arXiv:1408.5093*, 2014. 117

[176] C. N. Dickson, A. M. Wallace, M. Kitchin, and B. Connor, “Vehicle detection using multimodal imaging sensors from a moving platform,” in *SPIE Security+ Defence*. International Society for Optics and Photonics, 2012, p. 854112. 127

[177] J. A. Ratches, “Review of current aided/automatic target acquisition technology for military target acquisition tasks,” *Optical Engineering*, vol. 50, no. 7, p. 072001, 2011. 127

[178] H. Zhang, N. M. Nasrabadi, Y. Zhang, and T. S. Huang, “Multi-view automatic target recognition using joint sparse representation,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 48, no. 3, pp. 2481–2497, 2012. 128

[179] J. Sun, G. Fan, L. Yu, and X. Wu, “Concave-convex local binary features for automatic target recognition in infrared imagery,” *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, pp. 1–13, 2014. 128

[180] B. Russell, A. Torralba, C. Liu, R. Fergus, and W. T. Freeman, “Object recognition by scene alignment,” in *Advances in Neural Information Processing Systems*, 2007, pp. 1241–1248. 128

[181] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin, “Context-based vision system for place and object recognition,” in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 273–280. 128

[182] H. Izadinia, F. Sadeghi, and A. Farhadi, “Incorporating scene context and object layout into appearance modeling,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 232–239. 128

[183] P. Carbonetto, N. De Freitas, and K. Barnard, “A statistical model for general contextual object recognition,” in *European conference on computer vision*. Springer, 2004, pp. 350–362. 128

[184] C. Galleguillos, B. McFee, S. Belongie, and G. Lanckriet, “Multi-class object localization by combining local contextual interactions,” in *Computer Vision*

and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010, pp. 113–120. 128

[185] N. M. Robertson and J. Letham, "Contextual person detection in multi-modal outdoor surveillance," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European.* IEEE, 2012, pp. 1930–1934. 128, 132, 134, 135

[186] K. Murphy, A. Torralba, D. Eaton, and W. Freeman, "Object detection and localization using local and global features," pp. 382–400, 2006. 132

[187] A. Torralba, "Contextual priming for object detection," *International journal of computer vision*, vol. 53, no. 2, pp. 169–191, 2003. 132

[188] S. Gould, R. Fulton, and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," in *2009 IEEE 12th international conference on computer vision.* IEEE, 2009, pp. 1–8. 133, 135