University of KwaZulu-Natal
College of Agriculture, Engineering and Science

Candidate Generation and Validation Techniques for Pedestrian Detection in Thermal (Infrared) Surveillance Videos

by:

Oluwakorede Monica Oluyide

213554623

Supervisors:

Prof. Tom Walingo

Prof. Jules-Raymond Tapamo

in fulfillment of the academic requirements for the degree of Doctor of Philosophy in Computer Engineering, School of Engineering, University of KwaZulu-Natal

Submitted February, 2022

Abstract

Video surveillance systems have become prevalent. Factors responsible for this prevalence include, but are not limited to, rapid advancements in technology, reduction in the cost of surveillance systems and changes in user demand. Research in video surveillance is majorly driven by rising global security needs which in turn increase the demand for proactive systems which monitor persistently. Persistent monitoring is a challenge for most video surveillance systems because they depend on visible light cameras. Visible light cameras depend on the presence of external light and can easily be undermined by over-, under, or non-uniform illumination. Thermal infrared cameras have been considered as alternatives to visible light cameras because they measure the intensity of infrared energy emitted from objects and so can function persistently. Many methods put forward make use of methods developed for visible footage, but these tend to underperform in infrared images due to different characteristics of thermal footage compared to visible footage. This thesis aims to increase the accuracy of pedestrian detection in thermal infrared surveillance footage by incorporating strategies into existing frameworks used in visible image processing techniques for IR pedestrian detection without the need to initially assume a model for the image distribution. Therefore, two novel techniques for candidate generation were formulated. The first is an Entropy-based histogram modification algorithm that incorporates a strategy for energy loss to iteratively modify the histogram of an image for background elimination and pedestrian retention. The second is a Background Subtraction method featuring a strategy for building a reliable background image without needing to use the whole video frame. Furthermore, pedestrian detection involves simultaneously solving several sub-tasks while adapting each task with IR-specific adaptations. Therefore, a novel semi-supervised single model for pedestrian detection was formulated that eliminates the need for separate modules of candidate generation and validation by integrating region and boundary properties of the image with motion patterns such that all the fine-tuning and adjustment happens during energy

minimization. Performance evaluations have been performed on four publicly available benchmark surveillance datasets consisting of footage taken under a wide variety of weather conditions and taken from different perspectives.

Preface

This research discussed in this thesis was done at the University of KwaZulu-Natal, Durban from February 2016 to February 2022 by Oluwakorede Monica Oluyide under the supervision of Prof. Tom Walingo and Prof. Jules-Raymond Tapamo.

Declaration - Supervisor

As the candidate's supervisor, I agree to the submission of this thesis

---

Prof. Tom Walingo

Declaration - Co-Supervisor

As the candidate's co-supervisor, I agree to the submission of this thesis

---

Prof. Jules-Raymond Tapamo

Declaration - Plaigiarism

I, Oluwakorede Monica Oluyide, declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.
2. This thesis has not been submitted for any degree or examination at any other university.
3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.
4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
   (a) Their words have been re-written but the general information attributed to them has been referenced
   (b) Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.
5. This thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

Oluwakorede Monica Oluyide

Declaration - Publications

DETAILS OF CONTRIBUTION TO PUBLICATIONS that form part and/or include research presented in this thesis

Oluyide O.M., Tapamo JR., Walingo T. (2021) Fast Background Subtraction and Graph Cut for Thermal Pedestrian Detection. Lecture Notes in Computer Science, vol 12725, pp 219-228. 10.1007/978-3-030-77004-4_21

Oluyide O.M., Tapamo JR., Walingo T. (2022) Automatic Dynamic Range Adjustment for Pedestrian Detection in Thermal (Infrared) Surveillance Videos. *Sensors* 22(5):1728. 10.3390/s22051728

Oluyide O.M., Tapamo JR., Walingo T. (2021) Pedestrian Detection in Thermal Infrared videos using motion-constrained Graph-Cut. *PeerJ Computer Science* (Submitted)

Oluwakorede Monica Oluyide

Acknowledgments

I give glory to God for His faithfulness.

I am grateful to my supervisors, Prof. Tom Walingo and Prof. Jules-Raymond Tapamo for their commitment and dedication throughout this PhD journey.

I appreciate my family members for their care and support.

My sincere thanks go to my friends for their kindness to me.

Contents

List of Figures

List of Tables

Abbreviations

VSS	**V**ideo **S**urveillance **S**ystems
VGG	**V**isual **G**eometry **G**roup
EMR	**E**lectromagnetic **R**adiation
IR	**I**nfra**r**ed
NIR	**N**ear **I**nfra**r**ed
SWIR	**S**hortwave **I**nfra**r**ed
MWIR	**M**idwave **I**nfra**r**ed
LWIR	**L**ongwave **I**nfra**r**ed
TIR	**T**hermal **I**nfra**r**ed
HDR	**H**igh **D**ynamic **R**ange
FPA	**F**ocal **P**lane **A**rray
ROI	**R**egon **o**f **I**nterest
RGB	**R**ed **G**reen **B**lue
MCE	**M**inimum-**C**ross **E**ntropy
SNR	**S**ignal-to-**N**oise **R**atio
HOG	**H**istogram of **O**riented **G**radients
GBVS	**G**raph-**b**ased **V**isual **S**aliency
BMS	**B**oolean **M**ap-based **S**aliency
CNN	**C**onvolutional **N**eural **N**etwork
PiCANet	**Pi**xel-wise **C**ontextual **A**ttention **N**etwork
YOLO	**Y**ou **O**nly **L**ook **O**nce
ADAS	**A**dvanced **D**river **A**ssist **S**ystems
OSU	**O**hio **S**tate **U**niversity
LITIV	**L**aboratoire d'**I**nterprétation et de **T**raitement d'**I**mages et **V**idéo

TMIR	**T**erravic **M**otion **I**nfrared
LTIR	**L**inköping **T**hermal **I**nfrared
MCGCE	**M**otion-**c**onstrained **G**raph **C**ut **E**nergy
EHM	**E**ntropy-based **H**istogram **M**odification
BS2FI	**B**ackground **S**ubtraction **2** **F**rame **I**nitialisation
TTP	**T**otal **T**rue **P**ositive
OMT	**O**utdoor **M**otion and **T**racking
TFP	**T**otal **F**alse **P**ositive
PPV	**P**ositive **P**redictive **V**alue
TP	**T**rue **P**ositive
FP	**F**alse **P**ositive

Chapter 1

Introduction

Video surveillance has become a mainstay in Society. In 2016, the market for video surveillance was valued at $29.89 billion and is expected to increase to a value of $72.19 billion in 2022 [2]. Video Surveillance Systems (VSS) include a camera or group of networked cameras deployed in specific areas (see Fig. 1.1) to capture and send video data of activities and patterns of behaviours. Two main factors are driving the prevalence and widespread adoption of VSS.

The first factor concerns the increasing demand for security due to rising crimes rates and terrorism. Presently, it is not uncommon to find such cameras mounted in public places and private residences, but previously, they could only be found on government buildings or large organisations. The captured video data is monitored live for real-time proactive response or recorded and stored for future forensic investigation. The earliest systems consisted of simple recording, storage and playback facilities [4] which were Reactive systems useful for after-the-facts record of events. These were followed by Proactive systems useful for immediate response

FIGURE 1.1: Surveillance Cameras mounted in various locations [3]

to critical events as the events were monitored live by human operators. The current efforts are towards Preventive Systems to predict critical events before they happen using various cues deduced from real-time monitoring of video feeds. At the very least, they are expected to provide an alert for the possibility of a potentially disastrous event.

The ubiquity of VSS is also driven by a corresponding drop in the cost of acquiring and installing such systems due to technological advancement from analogue to digital surveillance. Most modern systems consist of IP cameras able to compress and store video which can be installed by the owner over a computer wired or wireless network, allowing for monitoring of video feeds at any time and from anywhere. These cameras can also be easily adjusted to meet the owner's changing needs, unlike older analogue VSS whose recording, storage and playback facilities were installed together in one location. The forecast for the market for IP-based video surveillance systems has been estimated to grow at a higher Compound Annual Growth rate (CAGR) during the period between 2019 and 2024 [5].

With the advancement in technology, intelligence and video analytics have been introduced into video surveillance making it an important research area. Human operators were solely responsible for video analysis; however, this is prone to missed events as human attention degenerates over time [6, 7] and it becomes increasingly expensive to employ a lot of operators to make up for this shortcoming. This, therefore, necessitates the inclusion of computer-automated techniques involving imaging and video processing algorithms to assist and/or take over from humans the task of video monitoring and analysis [8]. Also, the data generated from modern VSS have the potential to be integrated with data generated by the Internet of Things (IoT) thereby sharing data with connected devices and creating opportunities for better security solutions.

1.1 Motivation

Given the paradigm shift from reactive to proactive surveillance, a most desirable trait is for VSS to function in a persistent - all day, all night, every day of the week - manner. However, without or in poor light, surveillance cannot be carried out. The most prevalent cameras in VSS are the visible light cameras which work on the same principle as the human eye and share the same limitation, that is, they cannot 'see' if objects are not properly illuminated. Sensors in visible cameras collect visible light radiated off objects which are used to render images and video streams. It is the radiation from light that causes objects to be illuminated and thus detectable by the eye or the camera. At night, artificial lights need to be provided otherwise monitoring cannot take place and although there is illumination during the day, analysis of video stream can be hampered by over-illumination, uneven illumination and too many fine details captured in the image. Thus, it becomes necessary to augment the VSS with cameras that do not depend on illumination and can work all day and night.

A thermal camera does not need any artificial lighting whatsoever to work properly and it cannot be blinded by direct sunlight (see Figs. 1.2 and 1.3), and thus, thermal imaging cameras provide uninterrupted 24/7 surveillance regardless of the amount of light available [9]. Thermal cameras are useful as soon as it is possible to detect heat energy from objects. All objects, whether they are hot, at room temperature or frozen, emit energy from the infrared part of the electromagnetic spectrum called heat signature. Thermal cameras create images based on the amount of infrared detected from objects in the scene [10].

Thermal cameras have been around, but their size, complexity and cost meant that their use was only a practical option for the military, rescue services, security firms and such similar operations. They were used extensively by military operators under very low light conditions. Thermal cameras can see through smoke and fog (see Fig. 1.4) and they currently find utility in industrial, commercial, and consumer markets [10].

(a) Visible light camera (b) Thermal camera

FIGURE 1.2: Comparison of image clarity during night driving [11]

(a) Visible light camera (b) Thermal camera

FIGURE 1.3: Comparison of Visible light camera and Thermal image of a roadway in bright sunlight [11].

(a) Visible light camera (b) Thermal camera

FIGURE 1.4: Comparison of Visible light camera and Thermal image of a person behind smoke [12]

1.2 Problem Statement

Thermal images are not simple grayscale images, that is, they are not RGB images turned grayscale but radiometric data encoded using 14 to 16 bits. They are obtained from an array of detectors within the thermal cameras which collect infrared energy emitted by objects. This high resolution means that the camera can detect slight differences in infrared energy. And as surveillance scenes typically have areas with very different temperatures and objects with different levels of infrared emissivity, the resulting histograms have long tails. Thus, thermal images are properly referred to as high dynamic range (HDR) images. To display HDR images on 8-bit displays, they need to be tone-mapped. Tone-mapping is, essentially, rendering high dynamic range data to low-dynamic range displays. The problem is that target detection is highly dependent on tone-mapping. Without a target in mind, it is either the histogram of the tone-mapped image is similar to the long-tail histogram of the raw sensor data such that there is a lot of redundant information to work through or the resulting image is a global contrast-enhanced image that is visually pleasing, but which increases the difficultly level for target detection using computer algorithm as several objects get mapped to the same gray level because of reduced sensitivity. An example of an image with a long-tail histogram and its corresponding contrast-adjustment is shown in Fig. 1.5.

Though IR images are not affected by illumination problems of visible images, the appearance of objects varies depending on several other factors. Firstly, IR images display a measure of how much IR energy was emitted by objects in the scene and this amount depends largely on the emissivity and transmissivity of the objects and reflected IR from other objects in the vicinity [13]. This is an important point given the general assumption that IR images display the temperature of objects in the scene. The emissivity is a measure of how much IR is given off by the object itself. Transmissivity is a measure of how much IR can pass through an object. The measure of how an object reflects the IR from other objects can increase or decrease the amount of IR detected from it. Fig. 1.6) illustrates this point with

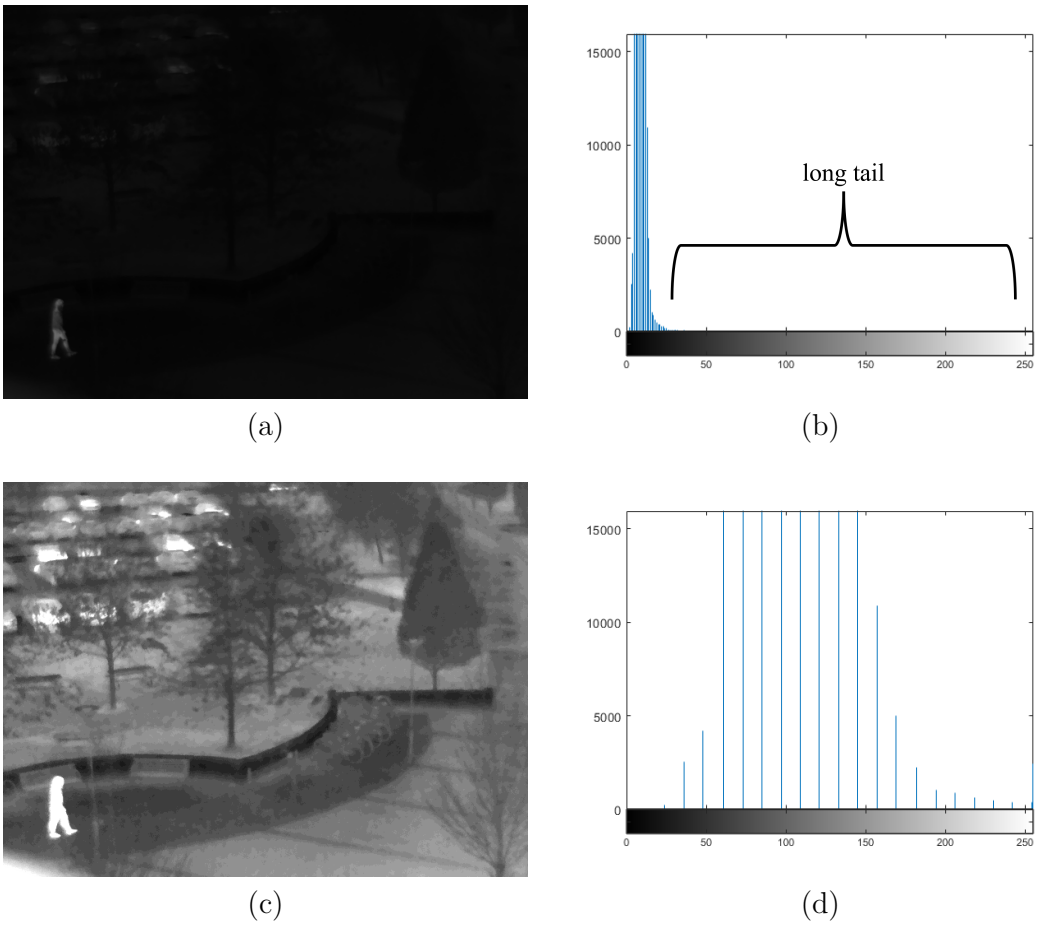
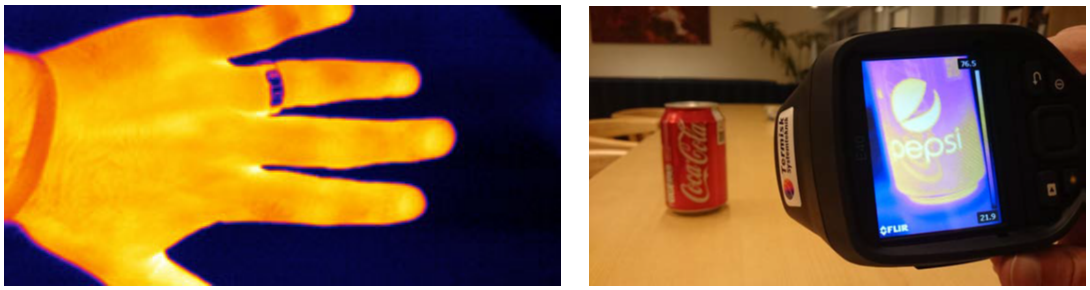


FIGURE 1.5: (a) IR image with (b) long tail histogram. (c) Same IR image but with (d) original long tail histogram clipped to the range of the high density gray levels and then stretched

an image of a hand with a ring on it. The ring and the hand are at the same temperature but are emitting IR at different rates. The ring appears to be much colder than the hand because it has a low emissivity and is reflecting a lot of the IR from the colder surrounding areas. This means that if the temperature of the environment were to change, the appearance of the ring will also change. Glass has a low emissivity and transmissivity and will almost always appear colder than its actual temperature.

Most footage taken for surveillance are captured in uncontrolled environments. Fig. 1.7 shows examples of surveillance footage from the same scene under different weather conditions. On some days, the white-hot appearance of the pedestrians shows very brightly, while on other days, there is polarity reversal in the footage



(a) (b)

FIGURE 1.6: Examples of how emissivity and reflected background can affect the perception of an infrared image.
(a) The ring appears to be much colder than the hand because it has a low emissivity and is reflecting a lot of the IR from the colder surrounding areas [14]
(b) A transparent tape with the Pepsi logo has been placed over the Coca-Cola Can filled with hot water. Because the tape has higher emissivity than the can, the tape appears warmer than the can which is why the details on the tape override those of the can [15]

and the pedestrians appear black. There are some days where the details in the background appear distinctly, while on other days they appear hazy.

(a) (b) (c)

FIGURE 1.7: Different weather conditions [16] (a) Morning, Partly cloudy, 25°C (b) Afternoon, Partly cloudy, 21°C (c) Morning, Haze, 18°C

Pedestrians are typically warmer than most objects in thermal video streams because the human skin emits (radiates) infrared energy almost perfectly emissivity at 0.98 where 1 is the value of the perfect radiator [13]. However, they are normally clad in accessories that reduce the amount of IR detected from major parts of the human body.

Certain limitations in Thermal Camera Technology present challenges to detecting Pedestrians. Generally, thermal cameras have lower spatial resolution and less sensitivity than visible cameras making it difficult to precisely define the shape or silhouette of the Pedestrian [18]. Infrared cameras using ferroelectric

Bird sequence [1]

Sequence 7 [16]

Sequence 7 [17]

Crouching Sequence [1]

FIGURE 1.8: Example of variation among IR images on images from different datasets. The Pedestrian's outfit mask a lot of the IR emitted from the human skin

detectors have unique limitations such as the presence of a halo around objects highly contrasting with the background.

The characteristics of thermal images influence the type of algorithm that can be used on them. It is generally found that the algorithms do not perform similarly on infrared and visible images for several reasons. For example, the statistical background technique used in [19] performs well on visible images but is hindered in performance by the artefacts present in thermal images ferroelectric detectors and an extra set of steps is required for ROI extraction. The effect of the artefact is shown in Fig. 1.9 where even after background subtraction, several steps are still required to extract the pedestrian from the halo. The performance of visible image algorithms are also hindered by the low amount of details present in infrared images. In recent times, the state-of-art object detectors have been applied to

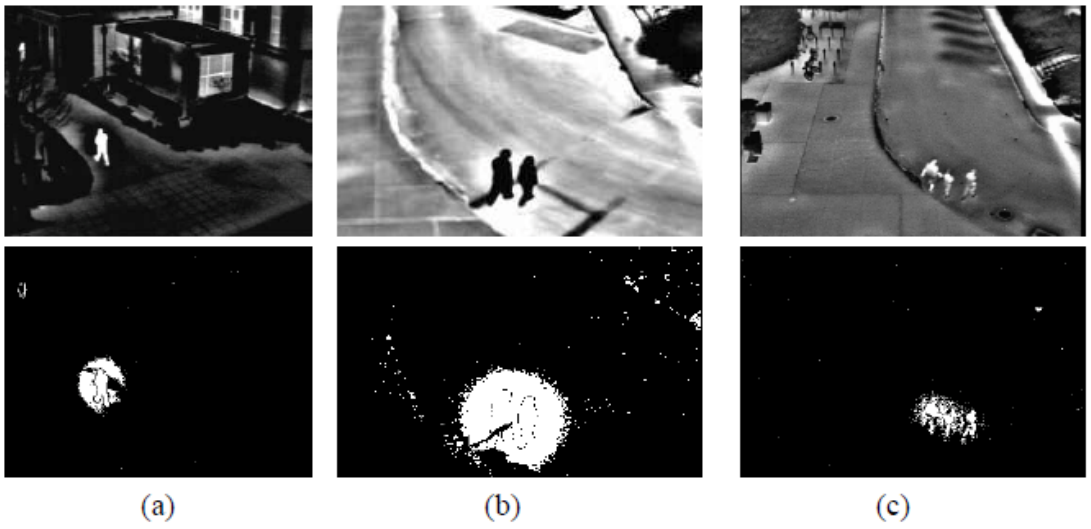


FIGURE 1.9: The performance of statistical background subtraction is hindered by the halo surrounding very cold or hold objects in infrared images as it detects both the pedestrians and the halo [19]

the task of pedestrian detection in infrared images. Some authors [20, 21] used YOLOv3 to detect pedestrians from infrared images in varying weather conditions. Gao et al. [22] performed for feature extraction using a transfer-learned visual geometry group (VGG-19) CNN. The rationale for using these methods is that they perform well on visible images and achieve state-of-the-art results. However, they do not perform similarly on infrared images for two reasons. First, the models developed for testing infrared images were trained on visible images. Second, different thermal cameras output different level of details. Also, even for models trained on infrared images such as done by [20], the performance of the trained model on different datasets will depend on the similarity of the test data to the training data.

1.3 Thesis Objectives

The aim of this research is to propose novel algorithms that adapt methods developed for visible images to infrared images to reduce the false-positive pedestrian detection rate. The specific objectives are:

1. To re-purpose histogram-based algorithms from contrast enhancement and unbiased image partitioning to background suppression for pedestrian ROI extraction

2. To formulate a background Initialisation method for Background Subtraction to extract moving regions

3. To create a single model for candidate generation and validation technique using Graph Cut

1.4 Thesis Contributions

The contributions of this thesis are as follows:

1. An entropy-based histogram modification algorithm for pedestrian hotspot detection. Histogram Modification, also referred to as Histogram Specification, is the transformation of an image's gray-level histogram to a desired histogram. The radiometric resolution of infrared images is quantified as the dynamic range of the image and is displayed in varying shades of gray which can be visualised with the aid of histograms. Normally, histogram modification is primarily done to enhance contrast for better visualisation by a human operator. However, in the proposed framework, histogram modification is done to extract likely pedestrian candidates. Given the two extremes of tone-mapping where there is either a lot of redundant information to work through or too little gray level variation to distinguish objects in the validation stage, the proposed algorithm seeks to reduce redundancies while leaving the original values intact for further processing.

2. A new Background Initialisation method for background subtraction which leverages motion. The background is modelled using only two video frames unlike most methods that require the use of the whole video. In addition,

it is robust against certain artefacts (depending on the sensor) in infrared images such as the halo effect which surrounds objects that are too hot or too cold and pedestrians which have different intensity levels in the same video frame.

3. Graph Cut is an optimization method for binary labelling problems that guarantees an exact solution. Graph Cut has not been used extensively in the thermal (infrared) domain for pedestrian detection. This study introduces Graph Cut as a validation method for Infrared Pedestrian detection.

4. A semi-supervised single-model for pedestrian detection that eliminates the need for separate modules of candidate generation and validation by integrating appearance properties of the image with motion patterns such that all the fine-tuning and adjustment happens during energy minimization. This model is inspired by the supervised single-model in [23] that integrates appearance and motion and the semi- supervised method in [24] that integrates the image region and boundary information into a single energy function. Both methods, [23] and [24], were used on visible images. The proposed framework shows significant improvement over using the original formulation in [24].

1.5 Organization of Thesis

The rest of the thesis is organized as follows.

- Chapter 2 presents background information on Infrared, thermal imaging and image analysis in the thermal domain and reviews the literature on pedestrian detection in thermal infrared images
- Chapter 3 presents the two Candidate Generation techniques formulated in this study
- Chapter 4 provides the experimental results and discussion

- Chapter 5 concludes the thesis and provides directions for future work

Chapter 2

Literature Review

2.1 Introduction

This chapter presents background information on Infrared, thermal imaging and image analysis in the thermal domain and reviews the literature on candidate generation and validation techniques for pedestrian detection in thermal infrared images

2.2 Infrared and Thermal Imaging

Not all light is visible. Light is essentially energy; another name for light is Electromagnetic Energy and the entire range of light is referred to as the Electromagnetic Spectrum (see Fig. 2.1). Electromagnetic energy travels in waves of different wavelengths typically measured in nanometres (nm). The eye easily detects the part of the spectrum called visible light which has wavelengths from 400-700 nm [25] and the skin can detect the wavelength 700 nm - 15,000 nm in the form of heat of the part of the spectrum called Infrared but the rest of the waves are imperceptible to the senses and can only be detected from their effects on an object or with the use of special instruments.

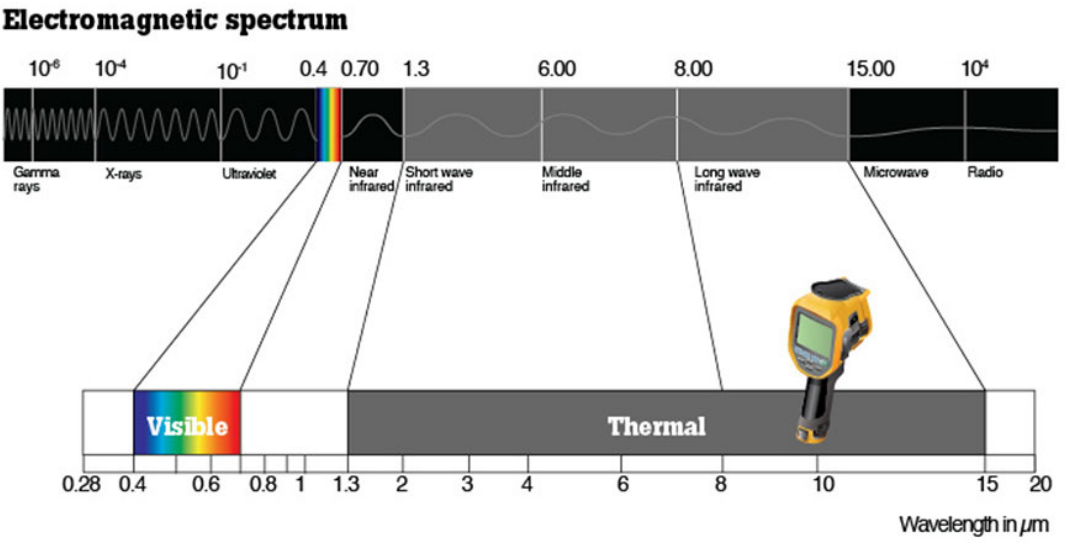


FIGURE 2.1: Diagram of the electromagnetic spectrum showing the different wavelengths of the various waves [14]

Sir Frederick William Herschel (1738-1822) discovered Infrared (IR) radiation in 1800 by its effects on a thermometer. The name "Infrared" is because it lies below the red light band of the visible light spectrum; in Latin, *Infra* means below. Red is a colour in the visible light spectrum which has the longest wavelength. Infrared energy has longer wavelengths than visible light with a range between 700nm and 30000nm. All objects emit infrared energy across this range of wavelength, but thermal sensors are sensitive only to a specific range [15].

IR can be subdivided into three bands based on the recommendation of the International Commission on Illumination (CIE) namely: IR-A, IR-B and IR-C.

TABLE 2.1: IR wavelength division based on CIE recommendation

Name	**Wavelength range μm**
IR-A	0.78 - 1.4
IR-B	1.4 - 3.0
IR-C	3.0 - 15

IR is also subdivided as follows: Near-infrared (NIR), Shortwave-infrared (SWIR), Midwave-infrared (MWIR), and Longwave-infrared (LWIR) and Far Infrared (FIR). Out of these divisions, LWIR and FIR constitute the part of the spectrum referred to as thermal infrared (TIR) [10, 15].

Table 2.2: Common IR wavelength division [10]

Name	CIE	Wavelength range	Usage
NIR	IR-A DIN	700nm - 1000nm	Fibre-Optic telecommunications
SWIR	IR-B DIN	1000nm - 3000nm	Long-distance telecommunications
MWIR	IR-C DIN	3000nm - 5000nm	Guided Missile technology
LWIR	IR-C DIN	8000nm - 14000nm	Thermal Imaging

Thermal images visually display measured amounts of thermal radiation emitted, reflected and transmitted within an area. The appearance of objects in the image is affected by multiple sources of thermal radiation, the object's properties and its surrounding. When Electromagnetic radiation (EMR) hits an object, some of it is absorbed, reflected and/or transmitted. Objects also embody thermal energy which can be converted into EMR. The amount of thermal energy emitted from an object depends on its temperature and the material it is made of. *Emissivity* is the ratio of the emittance of an object to the emittance of a black body. A black body, although it does not exist in nature, is defined as an object that absorbs all the incident EMR. The materials of an object exhibit properties different in the thermal spectrum than in the visual spectrum. For example, glass is opaque in the thermal spectrum but is transparent under visible light. Other materials that are reflective in one spectrum are usually not so in another. In addition, TIR cameras emit radiation during operation which changes the final appearance of the image.

Another challenge in thermal imaging is the type of thermal camera used and how the images are stored. Thermal cameras are of two types: cooled and uncooled. Cooled cameras are high-end cameras with high-temperature sensitivity that produce HD images often stored as 16-bit images to accommodate a wide dynamic range. Uncooled cameras are less expensive with lower temperature sensitivity and produce smaller noisier images. Thermal cameras calibrated to measure temperatures are called *Thermographic* cameras. *Radiometric* cameras allow access to raw 16-bit typically used to store images. All cooled thermal cameras are radiometric cameras, but not all uncooled cameras are. Most uncooled cameras prevent access to the raw 16-bit data and convert the images to 8-bit data thus adjusting the dynamic range of the image

to look visually pleasing to the eye but discarding a lot of information. While radiometric cameras' 16-bit images are desirable for our purposes, most of the images in public datasets have been stored as 8-bit images.

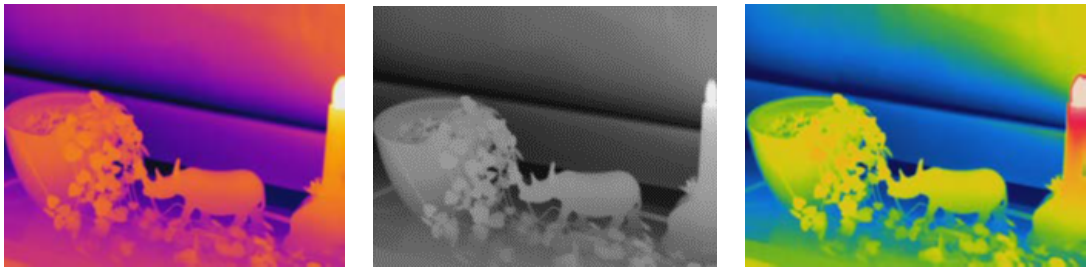
The major advantage of a thermal camera over a visible camera is that as thermal cameras measure emitted radiation and are independent of illumination, monitoring can be done at night under darkness and in difficult weather conditions. They are generally more advantageous in outdoor scenarios and at night. However, thermal cameras are more expensive, need more technical know-how for correct usage before operation and produce low-resolution images. In situations where there is no marked difference in thermal emissions and applications requiring identification of detected targets, visible cameras are more favourable.

2.3 Image analysis and Target Detection in TIR images

There are slight differences in image analysis when performed on visible and thermal images. The characteristics of TIR pose challenges and also nullify certain steps in algorithms for visible light image analysis. Some of the differences were introduced in Section 2.2.

In visible images, there is an immediate change in appearance as illumination changes while in thermal images, the appearance changes much slowly as emitted radiation gradually increases or decreases. Also, objects do not cast shadows in TIR images although emissions from highly reflective materials in the thermal spectrum like glass and water can appear similar to shadows. Therefore, for algorithms that depend on change such as background subtraction, the steps for scene update and shadow removal will not be as urgent in thermal imagery as in visible imagery.

Objects in the visible spectrum can easily be differentiated by their colour and are commonly displayed in the RGB (Red-Green-Blue) colour space. When visualised as histograms, three histograms provide information on how much red, green and blue are found in the scene. However, in the thermal spectrum, colours are useful only when they correspond to differences in temperatures or the property of materials in a scene. TIR images are commonly mapped to grayscale, but other colour maps exist also (see Fig. 2.2). The histogram of the TIR image corresponds to the amount of emitted radiation detected in the scene.



(a) Iron (b) Gray (c) Rainbow

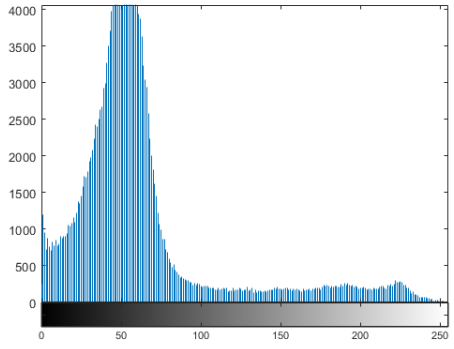
FIGURE 2.2: Visualising images under different colourmaps [15]

Although RGB images can be converted to grayscale images, they still do not show the same information as Infrared images. Fig. 2.3 shows the same scene captured using a visible camera and an infrared camera. The visible-light image has been converted to grayscale in Fig. 2.3 (b) and the histogram of the grayscale visible-light image and infrared images are shown in Fig. 2.3(d) and Fig. 2.3(e) respectively.

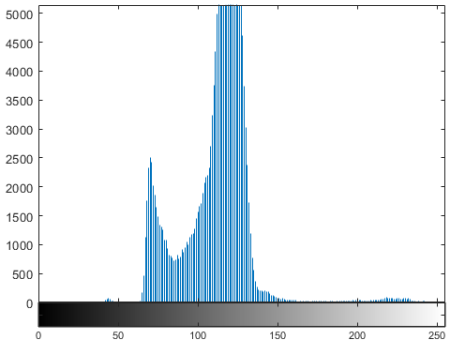
(a) RGB image

(b) RGB-turned-Grayscale image

(c) Infrared image



(d) Histogram of RGB-turned-Grayscale



(e) Histogram of Infrared image

FIGURE 2.3: Comparing visible image and infrared image [26] with their corresponding histograms. It can be observed that from the histograms that both images show disparate and disjoint information even though the visible colour image was converted to grayscale

2.4 Pedestrian Detection in Thermal Imaging

Pedestrian detection is one of the key tasks in the field of video surveillance as it provides necessary information for various activity tracking and behaviour understanding applications. Manual methods of detection are time-consuming and fraught with a high rate of missed detections while employing the necessary manpower to make up for the short attention span of humans is an expensive endeavour [6, 7]. Thus, efficient and accurate computer-automated methods in this regard are of high importance.

The pipeline for pedestrian detection in TIR images largely consists of steps to extract likely regions containing pedestrians and steps to discriminate between pedestrians and non-pedestrians within these regions Thus, it can be considered that there are two stages for pedestrian detection: Candidate Generation and Validation.

2.4.1 Candidate Generation

Candidate generation is directly responsible for the rate of true positive pedestrian detections. The two common techniques used in the thermal domain are thresholding and background subtraction while those based on saliency have also been found to be useful.

2.4.1.1 Thresholding techniques

Thresholding techniques rely directly on the differences in intensity values of the pixels in the image. Region-of-interest (ROI) extraction is carried out by applying a threshold to classify image pixels into those which belong to the target and those which belong to the background; one class of pixels will have intensities lower the threshold while the other class will have intensities equal to or higher than the threshold. Let $I(c, d)$ represent an image and $B(c, d)$ be the image after thresholding. $B(c, d)$ is defined as

$$B(c,d) = \begin{cases} 1, & \text{if } I(c,d) \geq T \\ 0, & \text{otherwise} \end{cases} \tag{2.1}$$

where T is the threshold value. Thresholding algorithms fall into two categories: parametric and non-parametric. Parametric algorithms obtain the threshold by parameter estimation. Non-parametric algorithms optimise an objective function for threshold selection. Both approaches may or may not assume or approximate distributions for the image histogram. Therefore, the innovation of methods in this category is either in the formulation of new threshold criteria to select an optimal threshold T and/or a new histogram approximation for the image.

Parametric methods produce excellent results when the assumed or approximated distribution of the histogram fits the image or dataset under consideration as the parameters are calculated based on the input image under study. Rajkumar and Mouli [27] proposed an algorithm consisting of steps to suppress the background and highlight the foreground pixels before employing local adaptive thresholding. The background is suppressed by subtracting the peak intensity of the histogram from each pixel value in the image whose value is greater than the peak intensity value. This process, in addition to suppressing the background, increases the image's signal-to-noise ratio (SNR). The foreground is highlighted by a special high-boost filter to improve the edges of the targets. The adaptive threshold is designed from the parameters (μ and σ) of the Gaussian distribution. An upper and lower limit threshold is calculated from the mean μ, the variance of the image σ together with an adaptive parameter k calculated using entropy. The pedestrian pixels are determined to be those that lie between the calculated upper and lower threshold limit. Let $B(c,d)$ be the image after adaptive thresholding. Its definition was given as

$$B(c,d) = \begin{cases} 0, & \text{if } \quad Th_l \quad \leq I_{mod}(c,d) \quad \leq Th_u \\ 1, & \text{otherwise} \end{cases} \tag{2.2}$$

where $I_{mod}(c,d)$ is the image after background suppression, $Th_l = \mu - k\sigma$ and $Th_u = \mu + k\sigma$ are the upper and lower threshold limits and $k = \sum_i^{L-1} p_f(i) \log_2 p_f(i)$.

Wu et al. [28] calculated the two threshold limits for binarisation by analysing the image histogram; a low threshold T_l to eliminate cold regions and a high threshold T_h to eliminate overly bright regions. They are given as follows

$$T_l = I_{peak} - \frac{I_{min}}{I_{mean}}\sigma \tag{2.3}$$

$$T_h = I_{peak} - \frac{I_{max}}{I_{mean}}\sigma \tag{2.4}$$

where I_{peak} is the peak pixel value of the distribution, I_{min} and I_{max} are the minimum and maximum pixel values of the frame under consideration respectively, I_{mean} is the mean and σ is the standard deviation of the frame pixel values. The binary image, $B(c,d)$ is given as follows

$$B(c,d) = \begin{cases} 255, & \text{if } T_l \leq I(c,d) \leq T_h \\ 0, & \text{otherwise} \end{cases} \tag{2.5}$$

where $I(c,d)$ is the input image.

Manda and Kim [29] put forward a formula for determining the optimal threshold for detecting objects in infrared images by approximating its histogram to the transient response of the first-order linear circuit. Following the observation that high-temperature regions correspond to foreground and object, they split the image histogram into two and model the second part of the histogram containing high gray level values as the transient response of the source-free first-order linear circuit given as

$$y = y(t_0)e^{-t(t-t_0)/\tau} \tag{2.6}$$

where $y(t_0)$ is the response of the system at initial time t_0. Therefore, the optimal threshold after histogram approximation was determined as

$$T = i_m + \tau \ln\left(\frac{h_m - h_{L-1}}{h_T - h_{L-1}}\right) \tag{2.7}$$

where i_m is similar to initial response time at t_0, h_m is similar to the response at $y(t_0)$ and h_{L-1} is similar to the response at the final time $y(t_1)$.

It is important to use the properties of the image to increase true positives. Parametric approaches can easily become are too image (dataset)-dependent. This is because it is difficult to generalise an algorithm when it has been specified for a particular set, but easier to specify an algorithm that has been created for a generalised purpose. Also, the parametric approach can be found to involve intensive computation and is not as accurate and robust as the non-parametric approach [30].

Non-parametric thresholding algorithms like [31–38] produce unbiased solutions useful in a wide array of applications and easily adaptable because the distribution of the histogram is not always explicitly assumed or approximated as in parametric approaches. Li et al. [30] find that the method by Otsu [31] and Hou et al. [33] do not perform satisfactorily on images whose object and background distributions are similar, as commonly found in thermal images. They propose a new criterion to separate an image with similar standard deviations into two parts making it more applicable to infrared images. The criterion was formulated as

$$J(t) = \min(\sigma_1(t), \sigma_2(t)) \tag{2.8}$$

and the optimal threshold determined as

$$T = arg \min_{t \in [0, L-1]} J(t) \tag{2.9}$$

Wu et al. [38] proposed a new criterion for finding the optimal threshold by modelling the object and background as normal cloud classes and constructing a criterion that compares the hyper-entropies of both cloud class models. The cloud model used assumes that the object and background distributions follow a normal distribution, and the hyper-entropy of each cloud model measures how much each class deviates from a normal distribution. The criterion aims to partition an image into two parts having high intraclass and low interclass similarities. The criteria based on the normal cloud model was formulated as

$$J(t) = \max(\mathrm{He}_b(t_1), \mathrm{He}_o(t_2)) \tag{2.10}$$

where $\mathrm{He}_b(t_1)$ and $\mathrm{He}_o(t_2)$ are the hyper-entropies for the background and object classes respectively and the optimal threshold was determined as

$$T = arg \min_{0<T\leq L-1} J(t) \tag{2.11}$$

Minimum Cross-Entropy algorithms [34–37] have drawn the attention of researchers because determining the threshold of images by comparing an image with its binary image produces promising results. Li and Lee [34] linearise the two images so that elements for comparison in both images are from the same location in the image, and they apply constraints to ensure that the distribution of the thresholded image follows closely that of the original image thus eliminating the need to assume the distribution of the histogram of the image. They formulated a criterion as follows

$$\eta(t) = -m_{1a}(t)\log(\mu_a(t)) - m_{1b}(t)\log(\mu_b(t)) \tag{2.12}$$

where $-m_{1a}(t)$ and $-m_{1b}(t)$ are the first moments of the object and background portions of a thresholded histogram with the range of gray level within $[1, L-1)]$ and $\mu_a(t)$ and $\mu_b(t)$ are the means for the object and background portions respectively. The optimal threshold was determined by

$$T = arg \min_{0<T\leq L-1} \delta(t) \tag{2.13}$$

Brink and Pendock [36] proposed the metric form of cross-entropy for threshold selection by imposing symmetry on the true (non-metric) cross-entropy formula to become

$$H_{CE}(T) = \left[\sum_{i=y}^{N} \mu_0(T) \log \frac{\mu_0(T)}{g(i)}\right]_{g_i \leq T} + \left[\sum_{i=1}^{N} \mu_1(T) \log \frac{\mu_1(T)}{g(i)}\right]_{g_i > T} \tag{2.14}$$

where $\mu_0(T)$ and $\mu_1(T)$ are the means of the above and below threshold portions of the image histogram and g is the distribution of the original grayscale image. In their work, the probability distribution is taken to be the normalised values of the histograms at each gray level.

Some minimum cross-Entropy algorithms for threshold selection assume a specific distribution for the image. Pal [35] assumes the data follows a mixture of Poisson Distribution. His work aimed to emphasise that when applying the minimum cross-entropy principles to images, the original image and the thresholded images must, indeed, be probability distributions and not simply a collection of gray levels, otherwise, the resulting algorithm cannot be called a "minimum cross-entropy algorithm for thresholding". The results, however, produced similar and sometimes poorer results to those in [34] due to the specific distribution chosen to model the image data.

The works reviewed so far using minimum cross-entropy were carried out on visible images. Minimum-cross entropy algorithms have not been utilised in TIR image analysis extensively and even less for pedestrian detection. However, it offers a great deal of flexibility that serves the purpose of this study which is to adapt visible light algorithms to IR images. One of its flexibilities is that it is not necessary to assume a specific distribution for the image histogram. This is

important for robustness because different conditions affect the appearance of IR images, and the distributions will be different corresponding to these conditions [39].

Lei et al. [40] use of maximum entropy for threshold selection in infrared images. Given the target entropy as H_{tar} and background entropy as H_{bkg}, let ϕ represent the sum of the target and background entropy as follows

$$\phi(th) = H_{tar} + H_{bkg} \tag{2.15}$$

Maximum information is obtained by maximizing $\phi(th)$ and the value of th at that point is the optimal threshold. After thresholding the image, they proceed to post-processing using morphology to eliminate noise and other obvious erroneous regions before obtaining the final extracted candidate regions.

2.4.1.2 Background Subtraction techniques

To reduce the dependence on intensity values for pedestrian detection, candidate regions are also generated by detecting moving regions. Optical Flow-based methods and Background Subtraction are commonly used for detecting moving objects in visible light images, and of the two, background subtraction is less computationally expensive [41].

The general procedure for Background Subtraction is to, first, create a model of the expected background, called background initialisation and then, to perform a comparison of the model with each frame in the video using a similarity function. This comparison is where the "subtraction" of background lies as the similarity function determines which pixels are labelled as likely pedestrian regions thus creating a Difference image. The innovation of methods in this category is usually in the background initialisation step of the technique.

Background initialisation is part of Background Subtraction that seeks to obtain an image that, when "subtracted" from each of the video frames, improves the

accuracy of detecting pedestrians in the scene. Ideally, the modelled background should contain no pedestrian so that any difference between the model and video frame is attributed to the pedestrian. Although many of the problems of background subtraction in visible imagery are nullified by the characteristics of IR images such as lighting changes and shadow removal, the problem of motionless pedestrians persists.

Jeon et al. [42] made use of temporal averaging over a sequence of images to obtain an initial background image that may still have pedestrians present. To ensure that pedestrians are totally eliminated, they proceed to find the humans present by applying a local max filter based on the consideration that pedestrians have a higher intensity value compared with the background. Next, they experimentally determine a threshold to binarise the image. With the humans detected, they apply a final procedure to erase them by linear interpolation with regions in the immediate vicinity. Although they obtain good results, there is still a high dependence on the intensity difference between pedestrian and background.

Jeyabharathi and Dejey [43] made use of frame differencing and reflected symmetrical pattern (RSP) to construct a background model. Frame differencing was used to detect the moving regions and the boundary of objects in the image. RSP provides geometrical information used in lieu of colour and gradient information. An object is said to possess reflectional symmetry when its reflection across a line appears as a mirror image. After frame differencing, the resulting image is divided into blocks and the RSP for each block is calculated.

Younsi et al. [44] made use of an adaptive Gaussian Mixture Model background subtraction method. The background distribution is chosen by sorting the distributions of the mixture model in descending order based on the ratio of the weight assigned to each distribution and its covariance matrix. Connected Component labelling is employed to improve the results of the background subtraction by discarding regions less than the size of the pedestrian in the scene.

2.4.1.3 Saliency-based Methods

Saliency maps are used to augment the information present in images. They guide the algorithm towards areas where more emphasis should be placed. This is done by increasing or decreasing feature weights. Candidate ROI generation using saliency algorithms can be grouped into two groups: bottom-up and top-down. Bottom-up approaches make use of the low-level features in the image relating to contrast such as colour, texture and orientation while top-down methods involve high-level features and prior knowledge.

State-of-the-art bottom-up saliency methods in visual imaging commonly used in IR imaging include those in [45–48]. Itti et al. [45] employ Difference of Gaussians (DoG) to determine centre-surround contrast regions. Harel et al. [46] put forward the Graph-based visual saliency (GBVS) which is an extension of Itti's method where normalisation is, instead, done using a graphical approach. Hou et al. [47] put forward the spectral residual (frequency domain processing) approach to saliency detection. Achanta et al. [48] put forward a frequency-tuned salient region detector that outputs a saliency map that is the same size as the image and with the detected salient regions having well-defined boundaries.

For pedestrian detection in IR imaging, Cai et al. [49] proposed a saliency mapping that fused local and global cues to overcome the shortcomings of the GBVS model [46] when used on IR images. The GBVS model calculates saliency based on local contrast while a spectral scale [47] is applied to calculate global saliency. The results of this method showed that pedestrians and vehicles have high saliency.

Lahouli et al. [50] put forward a hotspot ROI extraction algorithm for pedestrian detection that combines wavelet contrast enhancement with frequency-tuned saliency mapping [48]. The original image is first filtered with a Gaussian filter to eliminate high-frequency distortions before it is sent for contrast enhancement and saliency mapping. The output of the contrast enhancement and saliency detection modules are merged using their geometric mean and then binarised.

Yu et al [51] combined bottom-up and top-down methods in their framework. For the bottom up-model, they made use of the GBVS model [46] to extract ROIs based on the patterns of temperature where high-intensity values correspond to high temperatures. Unlike Itti's method and because they are dealing with IR images, they extract only two feature maps: luminance and orientation. For the top-down model, they introduced a gaze map to optimise the feature block selection used to create the HOG feature vectors.

Rajkumar and Mouli [52] employed an entropy and energy-based saliency mapping technique. The image is divided into block sizes of 15x15 and for each block, the local entropy *lent* and local energy *leng* are calculated. The global entropy *gent* and global energy *geng* are also calculated for the whole image and the saliency map is obtained as follows

$$I_{block} = \begin{cases} 1, & \text{if} \quad (lent > gent) \;\; \& \;\; (leng > geng) \\ 0, & \text{otherwise} \end{cases} \tag{2.16}$$

where I_{block} is the value assigned to each 15x15 image block.

Bottom-up approaches are more popular in the TIR domain than top-down approaches. This may be because top-down approaches are computationally expensive and inefficient for IR imaging [51]. Nonetheless, several authors [53–59] still make use of this approach to suppress background and highlight pedestrians features for training.

Heo et al. [53] and Truong and Kim [58] adapt the Boolean Map-based saliency (BMS) model from [60] for the pedestrian detection at night in IR images. Unlike the study with visible images in [60], only the luminance channel is enumerated as the feature channel in [53]. Also, the proposed algorithm took into consideration the changes in the contrast between the pedestrians and the road caused by changing weather conditions.

In [58], the feature channels are thresholded versions of the original image and the initial saliency map is a weighted sum of these versions. The final map is produced

by refining the initial map with thresholding, morphological area opening and Gaussian filtering. They create an improved version of the Otsu energy function for thresholding the IR image. The original Otsu objective function is given as

$$k^* = arg \max_{0 \leq k \leq 255} \{\phi_1(k)\mu_1^2(k) + \phi_2(k)\mu_2^2(k)\} \tag{2.17}$$

and the improved version which highlights small salient regions includes entropy ψ is given as

$$k^* = arg \max_{0 \leq k \leq 255} \{\psi(k)\phi_1(k)\mu_1^2(k) + \phi_2(k)\mu_2^2(k)\} \tag{2.18}$$

where k^* is the optimal threshold, ϕ_1 and ϕ_1 are the class occurrence probabilities and μ_1^2 and μ_2^2 are the mean class pixel values.

Chen and Shin [57] introduced an attention mechanism to an encoder-decoder convolutional neural network to re-weight features generated by the decoder module. Rahman et al. [59] make use of a saliency map generated by the deep saliency network (PiCA-net) in [55].

2.4.2 Candidate Validation

After the extraction of likely pedestrian candidate regions, the next step is to distinguish real pedestrians from similar regions. Current efforts make use of unsupervised or supervised approaches.

2.4.2.1 Unsupervised methods

Unsupervised candidate validation uses known or calculated physical properties of the pedestrians to separate them from non-pedestrians with similar properties.

The commonest property is the aspect ratio of the pedestrians which assumes that the humans are standing or in motion (walking, running, jogging). Caballero et al.

[61] combines the width to height ratio property with the standard deviation of the ROIs to distinguish pedestrians from incandescent regions. Pedestrians typically have higher standard deviation because heat distribution is not uniform across the entire body; uncovered parts of the body like the head tend to be brighter than covered parts. The authors also require that the containing area of pedestrians must be above a determined minimum area for discrimination.

Younsi et al. [44] proposed a global similarity function that uses the sum of sub-similarity functions to discriminate between human moving objects and non-human moving objects. A similarity threshold is established such that extracted regions of interest are classified as follows

$$\begin{cases} \text{“}Human\text{”}, & \text{if} \quad (S_{global} \geq Th_{sim}) \\ \text{“}Non-human\text{”}, & \text{otherwise} \end{cases} \tag{2.19}$$

where S_{global} is the sum of the sub-similarity functions and Th_{sim} is the similarity threshold. The sub-similarity functions include aspect ratio, foreground/background similarity ratio, overlapping similarity ratio, spatial distance between regions, and star-skeleton similarity function.

2.4.2.2 Supervised methods

The methodology for supervised validation includes feature extraction, training and testing. Whether feature extraction is a separate step in the methodology depends on the classifier framework. The most popular feature used in thermal imaging for pedestrian detection is the Histogram of Oriented Gradients (HOG) [62]. The use of these features corresponded with the use of Support Vector Machine (SVM) for classification [63]. Other features and/or classifiers pairs have been proposed. The innovation in this category is in the feature developed and in favourable classifier pairing. Although recent efforts are moving towards the use of convolutional neural networks (CNN) where features representation is

an inherent part of the training framework, feature representation is still a challenge for thermal imaging.

Wei et al. [64] combined HOG features with geometric characteristics of the training samples such as the ratio of bright pixel to total pixels, mean contrast and standard deviation and performed classification with Linear kernel SVM. Rajkumar and Mouli [65] combined HOG with Local Directional Pattern (LDP) and also performed classification with Linear kernel SVM. Baek et al. [66] developed a new HOG-based feature to handle the specific problems of pedestrian detection in TIR images as it incorporates not only thermal gradients but pixel intensities and their spatial locations. Classification was performed with Additive SVM. The computational cost of SVM increases as the number of candidates increases [40].

Li et al. [67] developed a mid-level feature to capture correlations among pixel-level features in addition to contrast and spatial pixel-level differences and pair this feature with a Principal Component Analysis (PCA)/SVM classifier. PCA is used for dimensionality reduction of the feature map which is then fed into the SVM classifier for discrimination. Their method was developed to overcome the problems of background subtraction in [16]. The performance depends on number of principal components chosen such that if the image is clear with good contrast, a larger number of principal components is required, and vice versa. Their future work involves looking for sparse features for thermal image representation.

Lei et al. [40] made use of Random Forest classifier with feature vector created using wavelet transform. Their method was compared with HOG+SVM [62] and showed improvements in the detection rate. They conclude the HOG is not a suitable feature for TIR because the images do not have the sufficient gradient information necessary for good performance of HOG.

Park et al. [68] proposed a CNN-based classifier with three input channels for fine-grained pedestrian detection. The input channels take in the original image, a Difference image from the previous frame and a background subtraction mask. They modified the output stride of the ResNet to accommodate pedestrians

occupying only small areas in the image and they used a weighted cross-entropy loss function to balance the skewed nature of IR image histograms where the background pixels are close to 99% of all the pixels in the image.

Chen et al. [57] developed an attention-guided autoencoder network that includes a skip-connection block. The function of this block is to combine features from the encoder-decoder modules such that contextual information is increased for robust and distinguishable features in IR images with low SNR and resolution. The method was able to detect pedestrians when the brightness values between the pedestrians and background were close but was limited by occlusions which, sometimes, resulted in false detections.

You Only Look Once (YOLO) is a deep learning algorithm that has achieved state-of-the-art as an object detector in visible images. Several studies [20, 53, 69–71] have been carried out on TIR images using various versions. Heo et al. [53] used tiny YOLOv2 for classification and adapted the framework to enhance pedestrians in IR images by adopting ABMS as the hand-crafted kernel. Huda et al [69] employed YOLOv3 to study the effect of transfer learning from visible images to the thermal domain to solve the problem of lack of large datasets for training deep models. My et al. [70] adapt a pre-trained YOLOv3 object detector to pedestrian detection in IR images by applying a generative data augmentation strategy. They use less than 50% real IR images for training and supplement with synthesised images. The synthesised images are created using a Least-squares Generative Adversarial Network. Li et al. [71] propose an improvement to YOLOv5 called YOLO-FIRI to include IR image features such that the network is forced to learn more discriminative features and small infrared object detection accuracy is improved.

2.5 Summary

This review presents several points for consideration. Firstly, Thermal Infrared (TIR) images are desirable in video surveillance because they can ”see” in

situations where visible light cameras cannot. Both modalities measure and display different quantities, therefore, a direct application of state-of-the-art algorithms that perform excellently on visible images will not achieve similar results on TIR images.

Secondly, Thermal cameras are useful as soon as there is a marked contrast between the target and the rest of the scene. Although this marked contrast, which refers to the difference in infrared emission level among objects in the scene expressed as pixel intensity when converted to digital images, can also be combined with motion and other geometrical properties for discrimination, it remains the foremost consideration. It can be seen that a lot of effort is being made to reduce the dependence on pixel intensity, contrast is always used as a boost, even in supervised methods.

Furthermore, algorithms for any particular task should have low computational requirements and be robust to images acquired under different conditions and different devices, but the balance of these two requirements must be determined by the level of accuracy achieved. Some authors choose low-cost computation techniques applicable to one public dataset and/or one private dataset with high accuracy while other authors choose high computation cost techniques applicable to several datasets with lower accuracy. As will be observed in Chapter 4 when the results are presented, detection accuracy reduces as the computational complexity of the algorithm increases. However, it is difficult to say whether this is a negative or positive outcome because algorithms at the two ends of the spectrum (low and high computational cost) generally do not use the same dataset for performance evaluation, and also, because the performance of models across different datasets depends on the similarity to the training data [20]. The question is, "given the nature of infrared images and the differences in image quality and sensitivity of thermal cameras, how much robustness is required from a proposed algorithm?", where robustness refers to high accuracy across different datasets. The answer to this question may lay in defining the limits of the study.

In this study, the aim is to detect pedestrians in surveillance videos whenever the output of thermal infrared cameras has more advantage over those from visible cameras. This means that thermal cameras are only considered as substitutes for visible cameras and in extreme weather conditions where the pedestrian is indistinguishable from the road or the rest of the background, the recommendation will be to use data from visible images or fuse them with thermal data for pedestrian detection. Also, applications in advanced driver assist systems (ADAS) and autonomous driving are excluded from the scope. Within this limit, robustness will be shown based on accuracy despite advancements in thermal camera technology under different acquisition conditions. With regards to computation time, the aim is to find a balance between low and high computation cost algorithms without losing accuracy.

The proposed methods formulated in this research will be presented in Chapter 3. In the proposed methods for candidate region generation, histogram distribution is not assumed or approximated to increase applicability across TIR pedestrian detection databases for surveillance. For the semi-supervised candidate validation where a model is assumed for each class, the information for building each class model will be user-specified to ensure data dependence.

Chapter 3

Materials and Methods

3.1 Introduction

Details of the candidate generation and validation techniques for pedestrian detection in Thermal Infrared (TIR) images formulated in this research are presented in this chapter. The first candidate generation technique proposed is an Entropy-based histogram modification algorithm and the second is a Background subtraction method featuring a 2-Frame Background Initialisation algorithm. The candidate validation technique presented is a motion-constrained Graph Cut -based framework. Preceding these sections will be details of the dataset used in this research.

3.2 Dataset

The TIR surveillance data used in this research were obtained from four publicly available databases namely, Ohio State University (OSU) thermal pedestrian database [16], LITIV (laboratoire d'Interprétation et de Traitement d'Images et Vidéo)) dataset [17], Terravic Motion IR (TMIR) database [72] and the Linköping Thermal InfraRed (LTIR) dataset [1]. These databases were selected many research efforts for pedestrian detection in IR have used them for

evaluation and to showcase the performance of the proposed methods on older and modern thermal cameras.

Ohio State University (OSU) thermal pedestrian database contains ten sessions of 360 x 240 thermal images culminating in a total of 284 frames each having an average of 3-4 people. A Raytheon 300D thermal sensor with a 75mm lens camera mounted on an eight-storey building was used to capture the scene. Each video sequence contains a comprehensive description of the weather conditions under which they were acquired. Ground truth is also available in the form of bounding boxes for each pedestrian detected in the scene.

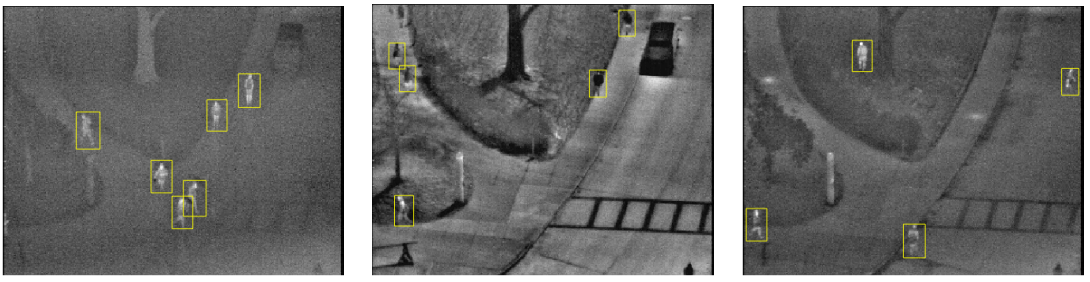


FIGURE 3.1: Sample images from OSU database showing the ground-truth bounding boxes

LITIV dataset contains 9 sequences of 320 x 240 thermal videos captured at 30 frames per second with different zoom settings from relatively high altitudes and at different positions culminating in a total of 6325 frames of lengths varying between 11s and 88s. Ground-truth was provided in the form of foreground binary masks. Sample images are shown in Fig. 3.1.

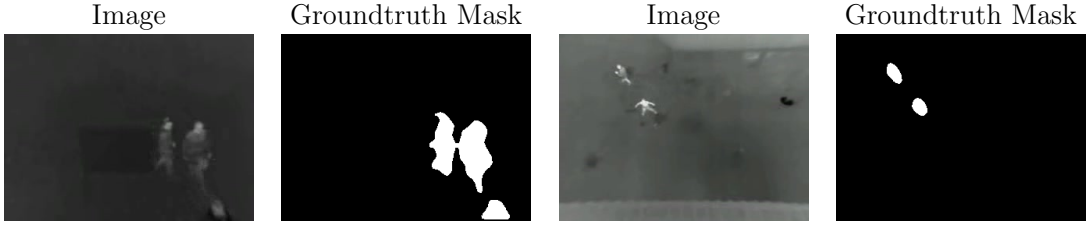


FIGURE 3.2: Samples from LITIV database showing the image and ground-truth binary mask

TMIR database features 18 thermal sequences with 8-bit grayscale JPEG images of size 320 x 240 pixels taken with a Raytheon L-3 Thermal Eye. Eleven sequences

were chosen from the Outdoor Motion and Tracking (OMT) Scenarios. Ground truth was not provided for this database. Sample images are shown in Fig. 3.3.

FIGURE 3.3: Sample images from TMIR database

The LTIR dataset consists of 20 thermal IR sequences featured in the Visual Object Recognition challenge 2015. Sequences pertaining to pedestrian detection were chosen. Each sequence in the database features a different scene. Groundtruth was provided in the form of coordinates. Sample images are shown in Fig. 3.4.

FIGURE 3.4: Sample images from LTIR database

3.3 Entropy-based histogram modification algorithm

This section presents a novel histogram specification algorithm based on entropy for pedestrian candidate generation in infrared images. In simple terms, Histogram specification is used to transform or modify a histogram to a desired or "specified" shape. However, rather than a transform, this research presents a strategy for achieving the desired goal but unknown shape. Histogram modification is used interchangeably with Histogram Specification.

By considering an image as a random variable m with a probability distribution function $P_m(w)$ specifying the probability that the value of the random variable

will be less than or equal to w, a transformation is function T that maps m to a new random variable n having a probability distribution of $P_{T \cdot m}(w)$. The task of histogram specification algorithms in this regard is to find a transformation T such that $T \cdot m$ has a desired distribution $P_d(w)$. Therefore, given the original image m and the desired histogram $P_d(w)$, the task is to find a T such that $P_{T \cdot m}(w)$ is metrically (or in some other sense) similar to $P_d(w)$. The problem is that it is difficult to predict the desired histogram because the exact range within the histogram to localise the pedestrians is not known; there is only a general idea that pedestrians are typically warmer than most objects in thermal video streams because the human skin emits (radiates) infrared energy at an almost perfect emissivity rate of 0.98 where 1 is the value of the perfect radiator [13]. Therefore, the information being sought is toward the right side of the histogram. The question is "how much to the right?" and "how will this movement to the right be guided?". Thus, rather than a transformation T, a strategy is used for increasing information loss that transforms the histogram of the original image into that of our desired goal and outputs an image with likely pedestrian candidate regions and minimal background.

Generally, there is loss of information when transforming m by T which means that constraints are usually included to compromise between the degree of similarity between $P_{T \cdot m}(w)$ and to $P_d(w)$ and the loss of information. While information loss is generally undesirable, it is desirable in this study because it simulates the effect of reducing the sensitivity of the thermal sensors which reduces details in the infrared images. Sensitivity reduction enables easier pedestrian region detection. As mentioned earlier, the radiometric resolution of the infrared images is dependent on the sensitivity of the thermal sensors and is quantified as dynamic range. With respect to thermal imaging, the dynamic range of a scene is the ratio of the hottest to the coldest region in the scene. The dynamic range of the camera is the ratio of the highest to lowest temperature the Focal Plane Array (FPA) or thermal sensors are capable of detecting. Sensitivity reduction is achieved by reducing the width between the highest and lowest infrared emission the FPA or thermal sensors is capable of detecting so that a

narrower dynamic range from the camera is mapped to a wider dynamic range on the display. The effect of imitating sensitivity reduction is shown in Fig. 3.5. It can be seen that by finding the range of intensities within which the hand lies and remapping that range to a wider range, pixels of the accessories on the hand (ring and hand-band) become saturated leading to fewer details. Saturation means that information is lost for pixels lower or higher than the minimum and maximum intensity range within which the hand lies.

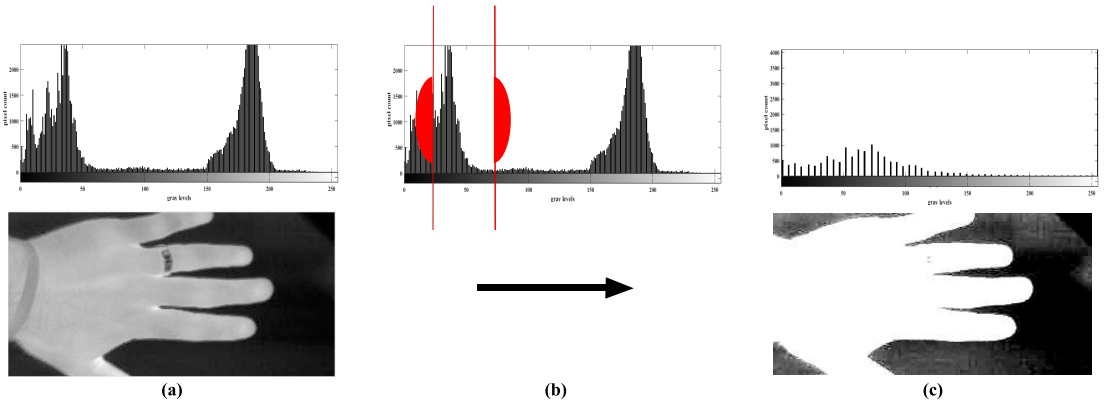


FIGURE 3.5: Manual sensitivity reduction forces pixels below and above the minimum and maximum values of the manually selected range for the hand to become saturated thereby reducing the details within the hand (a) original image and histogram (b) a chosen range within which the hand lies (best viewed in colour) (c) adjusted image and readjusted histogram

While Fig. 3.5 was done manually, the proposed algorithm described in this section aims to do it automatically by combining histogram equalisation and minimum cross-entropy for threshold selection for the new dynamic range. The success of the method lies in the fact that Entropy is more when information is uniformly distributed. By increasing the entropy within the human regions, that is sensitivity reduction, the areas of lower entropy are iteratively eliminated until the desired

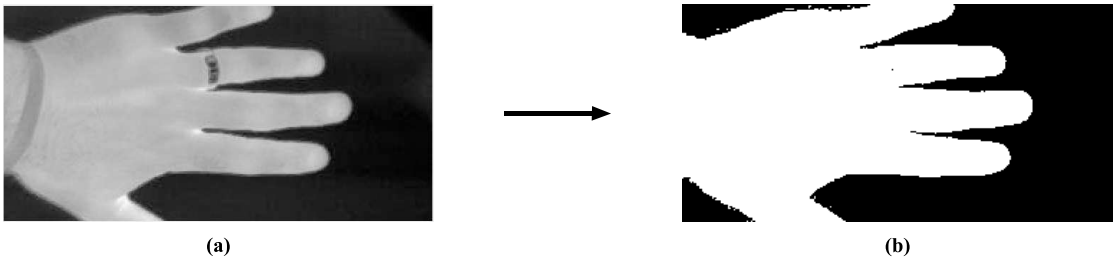


FIGURE 3.6: The result of using the proposed method on the hand image. (a) Input image (b) Output image

goal is attained. Histogram equalisation is responsible for increasing the entropy within the human regions while minimum cross-entropy provides the criteria for subsequent histogram modification until convergence is reached. Applying the proposed method on the image in Fig. 3.5 produces the result shown in Fig. 3.6.

An outline of the proposed framework is shown in Fig. 3.7 and is also based on the principle of minimum cross-entropy but it is unlike previous methods [34–37, 40, 73] in three ways.

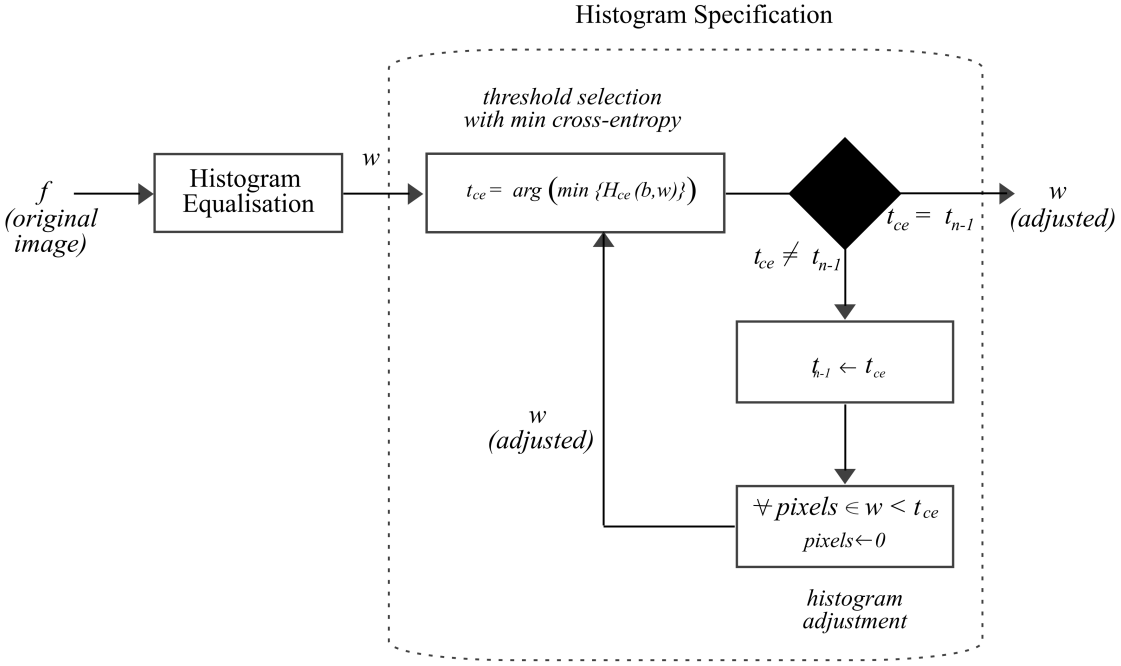


FIGURE 3.7: Overview of the proposed Entropy-based histogram modification method

First, rather than using the image in its original form, the image's histogram is equalised to become uniformly distributed. This is an important step because entropy increases when information is uniformly distributed and as the proposed strategy involves information loss, it is necessary to maximise, as much as possible, the entropy of the areas to be kept. Also, it was found that equalising the image provides a spatial cohesiveness to the infrared image which is necessary for the success of the iterative nature of the proposed algorithm. Secondly, given the wide dynamic range of infrared image, the sensitivity of the image was iteratively reduced. Furthermore, the algorithm is "guided" towards the right side of the histogram by making all the pixels intensities in the

equalised image less than the intensity at the minimum cross-entropy t_{CE} equal to zero. Therefore, at each iteration, the gray level at which minimum cross-entropy is attained becomes the new minimum dynamic range value for the image thus creating a modified histogram for the next iteration until the previous and current minimum value become equal. Thirdly, several algorithms rely on normalizing the pixel intensity to provide a measure of the probability, but histogram equalisation achieves the same effect. Therefore, the known probabilities $w(x_i)$ are the equalised pixel intensity values of the original image.

3.3.1 Histogram Equalisation

The mathematical foundation of histogram equalisation is based on the idea that pixels in the original and equalised images can be regarded respectively as continuous random variables H and N in the range of graylevels $[0, L-1]$ and the normalized histogram as probability density function (PDF) [74]. It is a transform T of H into N which spreads gray levels over the entire scale and each gray level is allotted an equal number of pixels. Therefore, Y is defined as

$$N = T(H) = (L-1)\int_0^H p_H(h)dh, \tag{3.1}$$

where p_H is the PDF of the original image. T, therefore, is the cumulative distribution of H multiplied by $(L-1)$.

The histogram h_f of an image f with L gray levels in the range [0, 255] is given as

$$h_f(l) = n_l \tag{3.2}$$

where n_l is the number of pixels with l graylevel. If the image has m pixels in total, the normalised histogram p_f is calculated as

$$p_f(l) = \frac{h_f(l)}{m} \tag{3.3}$$

where $l = 0, 1, ..., L-1$. Given the graylevel l, this is equivalent to dividing each gray level n_l by the total number of pixels in the image m. Given the intensity k, the histogram equalised image w of f can then be defined by

$$w(k) = \text{floor}((L-1)\sum_{l=0}^{k} p_f(l) \tag{3.4}$$

3.3.2 Histogram Specification

3.3.2.1 Cross-Entropy

Let the original image be the prior distribution with known probabilities $w(x_i)$. Cross-entropy $H_{CE}(b, w)$ is the average number of bits needed to encode data with distribution $b(x_i)$ when modelled with distribution $w(x_i)$ [75] and is defined as

$$H_{CE}(b, w) \ = \sum_i b(x_i) \log \frac{1}{w(x_i)} \ = -\sum_i b(x_i) \log w(x_i)$$

and can be written as

$$H_{CE}(b, w) = H_E(b) + D_{KL}(b||w) \tag{3.5}$$

where

$$H_E(b) = -\sum_i b(x_i) \log b(x_i)$$

is the entropy of the distribution b and

$$D_{KL}(b||w) = \sum_i b(x_i) \log \frac{b(x_i)}{w(x_i)}$$

is the Kullback-Leibler (K-L) divergence of distributions b and w defined as the excess code over the optimal code needed to represent data because it was modelled

using distribution w instead of the true distribution b [75]. The value of $H_E(b)$ in equation (3.5) is fixed and during minimization reduces to an additive constant. Therefore, the cross-entropy reduces to

$$H_{CE}(b, w) = \sum_i b(x_i) \log \frac{b(x_i)}{w(x_i)} \tag{3.6}$$

subject to

$$\sum_i b(w_i) = \sum_i w(x_i) \ (= 1)$$

3.3.2.2 Minimum Cross-Entropy for minimum range value selection

The minimum range value selection problem can be posed as the choice of the best distribution estimate for an event with unknown probabilities. Let the event with unknown probabilities $b^+(x_i)$ be the modified image where x_i refers to the graylevel of the image pixels or the number of bins in the image histogram. The problem is to choose a distribution b that best estimates b^+ given what is known. The solution to the problem is the distribution having expected values that fall within the bounds or equal to the known values thereby satisfying certain learned expectations $\sum_i b^+(x_i)m_d(x_i)$ or constraints. However, there is an infinite set of distributions that satisfy the constraints. Information is a measure of one's freedom of choice in making selections [76] and thus, entropy, a measure of information, becomes necessary. The principle of maximum entropy, which states that the distribution of choice, from all that satisfy the constraints, is the one with the largest entropy is the prescribed solution for solving such problems. However, in situations where a prior distribution that estimates b^+ is known in addition to the learned expectations, the principle of minimum cross-entropy, a generalisation of the maximum entropy principle, applies. The principle of minimum cross-entropy states that of all the distributions b that satisfy the constraints, the distribution of choice is the one with the smallest cross-entropy [77].

The unknown probability distribution b is determined from that of the equalised original image w and can be described using two probability measures $\mu_0(T)$ and $\mu_1(T)$ which are the below- and above-threshold means of the equalised original image respectively. These expectations are summarised as

$$\begin{aligned} (i)\ \ & b(x_i) \in \{\mu_0(T), \mu_1(T)\} \\ (ii)\ \ & \sum_{w(x_i)<T} w(x_i) = \sum_{i<T} \mu_0(T) \\ (iii)\ \ & \sum_{w(x_i)\geq T} w(x_i) = \sum_{i\geq T} \mu_1(T) \end{aligned} \tag{3.7}$$

which allows for the determination of $\mu_0(T)$ and $\mu_1(T)$ as

$$\begin{aligned} \mu_0(T) &= \sum_{i=y}^{T} ip_i \\ \mu_1(T) &= \sum_{i=T+1}^{z} ip_i \end{aligned} \tag{3.8}$$

where y and z are the lowest and highest graylevels present in the equalised original image respectively, T is the candidate threshold value and p_i is the probability of gray level i given by:

$$P_i = \frac{m_i}{N} \tag{3.9}$$

where m_i is the number of pixels having graylevel x_i and N is the total number of pixels making up the image. Therefore,

$$\sum_{i=y}^{i=z} w(x_i) = \sum_{y}^{i<T} \mu_0(T) + \sum_{i\geq T}^{z} \mu_1(T) \tag{3.10}$$

Hence, the Cross Entropy from equation (3.6) becomes

$$H_{CE}(b,w) = \sum_{i=y}^{T} \mu_0(T) \log\left(\frac{\mu_0(T)}{w(x_i)}\right) + \sum_{i\geq T}^{z} \mu_1(T) \log\left(\frac{\mu_1(T)}{w(x_i)}\right) \tag{3.11}$$

and the pixel intensity value which chosen corresponds to the minimum Cross Entropy and is given as

$$t_{CE} = arg\left(\min_{y<T\leq z}\{H_{CE}(b,w)\}\right) \tag{3.12}$$

which is the minimum range value.

3.3.2.3 Histogram Adjustment

Let t_{n-1} be the previous t_{CE} and t_n be the current t_{CE}. The decision on whether the distribution b is the final distribution is determined by if $t_{n-1} = t_n$ where t_{n-1} and t_n is the previous and current t_{CE} respectively. Consider an image f with histogram h_f with L gray levels and $l = 0, 1, 2, ..., L-1$. After each iteration, if $t_{n-1} \neq t_n$, all $l < t_{CE}$ are set to zero which readjusts the dynamic range of the image. Thus, a new t_{CE}, referred to as t_n for the sake of clarity, can be calculated as

$$t_n = arg\left(\min_{t_{n-1}<T\leq z}\{H_{CE}(b,w)\}\right) \tag{3.13}$$

and compared to t_{n-1}. This means that at every iteration, a different histogram-adjusted image takes the place of the original w in Eq. 3.12 and the minimum range value using minimum cross-entropy is re-calculated. If there is no change, that is $t_{n-1} = t_n$, the iteration stops and the resulting w is chosen as the image containing the likely candidates.

Figs. 3.8, 3.9, 3.10 and 3.11 show the modified histograms at each iteration for different images from three different databases. These are the images which take the place of the w in Eq. 3.12 at each iteration.

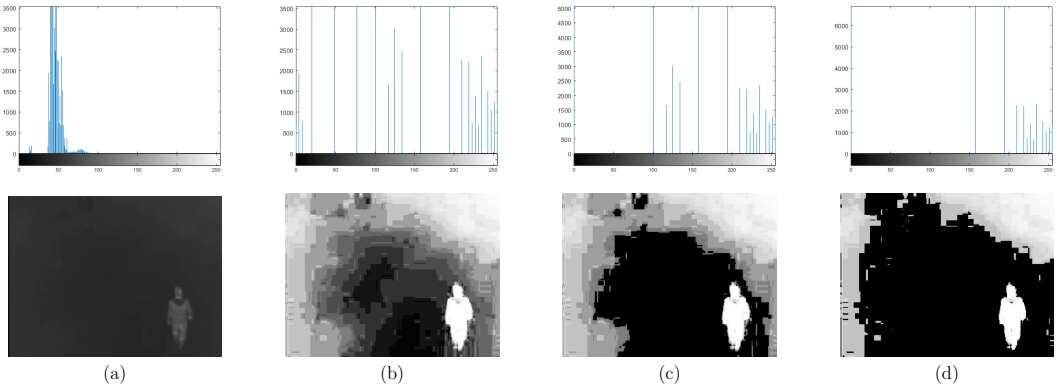


FIGURE 3.8: Histogram Adjustment (image in000399 (Sequence 3) from LITIV database)

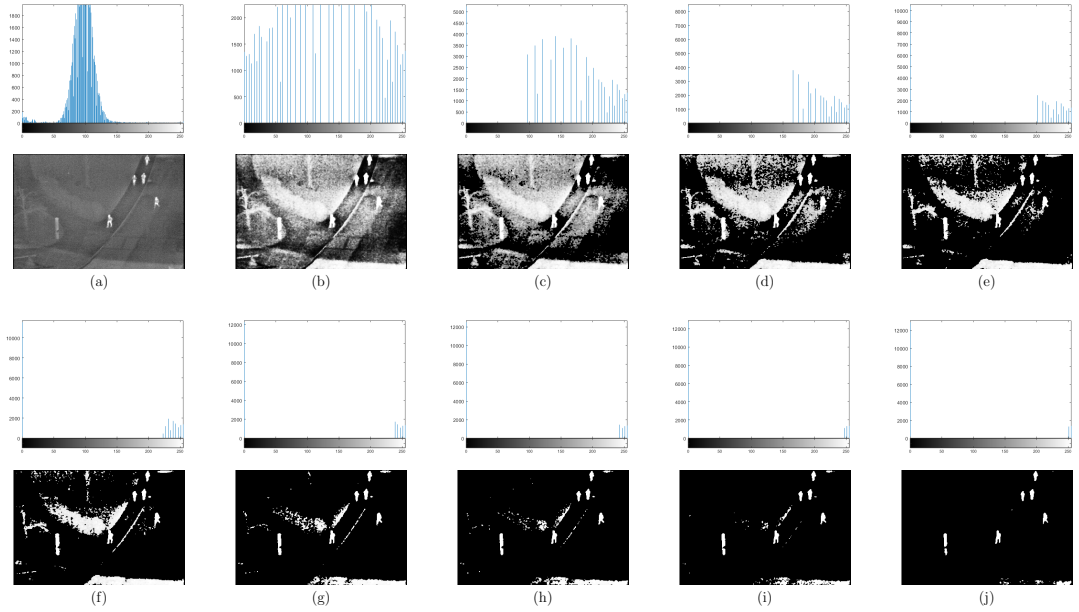


FIGURE 3.9: Histogram Adjustment (image img_00014 (00002) from OSU database)

The procedure of the proposed framework is also summarised in Algorithm 1

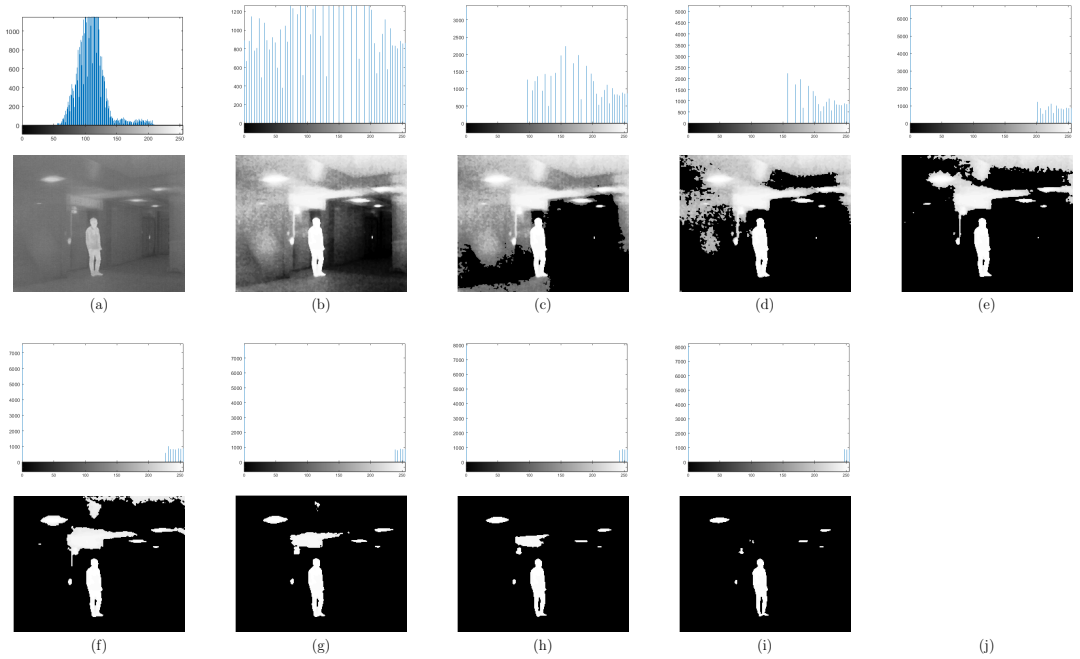


FIGURE 3.10: Histogram Adjustment (image 00000006 (hiding) from LTIR database)

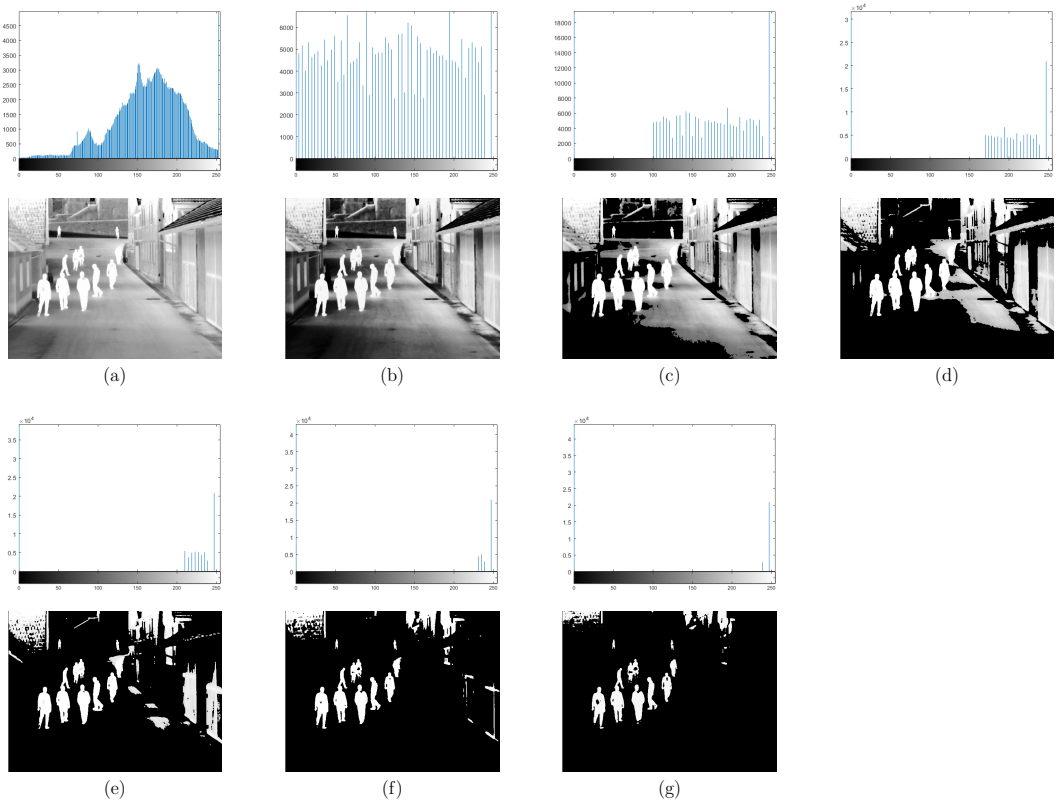


FIGURE 3.11: Histogram Adjustment (image 00000005 (Saturated) from LTIR database)

Algorithm 1 Entropy-based histogram modification

Input: Input image **f**,
Output: Adjusted image **w** with pixels $i \in \mathbf{w}$
obtain **w** using Eq. 3.4
calculate t_{CE} from **w** using Eq. 3.12
$t_n \leftarrow t_{CE}$; $t_{n-1} \leftarrow 0$
for $i \in \mathbf{w}$ **do**
    if $i < t_{CE}$ **then**
        $i \leftarrow 0$ *// the output is a new* $\boldsymbol{w}$
    end if
end for
while $t_n \neq t_{n-1}$ **do**
    $t_{n-1} \leftarrow t_{CE}$
    Compute t_{CE} on **w** using Eq. 3.13
    for $i \in \mathbf{w}$ **do**
        if $i < t_{CE}$ **then**
            $i \leftarrow 0$
        end if
    end for
    $t_n \leftarrow t_{CE}$
    // the output is a new $\boldsymbol{w}$ *until the while condition is false*
end while

3.4 Background Subtraction using 2-Frame Background Initialisation

This section presents a novel background subtraction algorithm that features a 2-frame background initialisation algorithm for candidate generation in Thermal Infrared images. Background initialisation builds a representation of the scene without the targets. This algorithm proposes to create a useful background image from two frames in the whole video.

This algorithm is based on the nature of thermal infrared (TIR) images. Firstly, TIR footages do not change as rapidly as their visible counterparts. This means that scenes do not suddenly become bright or become dark because, for example, the sun suddenly came out or was hidden by a cloud. Also, the appearance of objects does not change rapidly as the increase or reduction of thermal emissions happens gradually. For example, a warm pedestrian standing in the rain will gradually cool down over time beginning with the feet and hands. Secondly, some

unwanted regions remain even after utilizing the whole video information. For example, the work in [42] still had to detect pedestrians as having higher intensities to totally eliminate them after averaging all the frames in the sequence. Also, some artefacts such as halos become aggravated such as in [19] when the whole sequence is used. Thirdly, the low resolution and noisy nature of infrared images mean that only slight modification in the pixel values where motion is detected is necessary to obtain a background able to detect pedestrians in the rest of the video frames.

Two frames are required because motion is used as the criteria for extracting the regions of interest. Also, two frames are used to keep the algorithm computationally inexpensive. Motion foster the usefulness of this algorithm in a wider variety of images where hotspot algorithms fail such as reverse polarity and pedestrians with non-uniform appearance.

An outline of the proposed background subtraction method is presented in Fig. 3.7. Let $(M_g)_{0\leq g\leq U_{k-1}}$ represent a sequence of U_k video frames. First, a Difference $\mathcal{D}_f$ is obtained as follows

$$\mathcal{D}_k = |M_k - M_{k+step}| \tag{3.14}$$

where *step* in M_{k+step} refers to the number of frames between the two selected frames.

Let $G + H$ be the pixel-wise sum of pixels values of two images G and H, and G^c represent the inversion of each pixel value of an image G. $G + H$ is the "distortion of pixel where motion is detected" mentioned at the beginning of this section. Thus, the background image $\mathcal{B}$ is obtained as

$$\mathcal{B} = (\mathcal{D}_k + M_k^c)^c \tag{3.15}$$

Background subtraction is then carried out to generate likely candidate regions. For each frame M_p, the likely candidates are generated as follows

$$\mathcal{LC}_p = |\mathcal{B} - M_p| \quad 0 \leq p < U_{k-1} \tag{3.16}$$

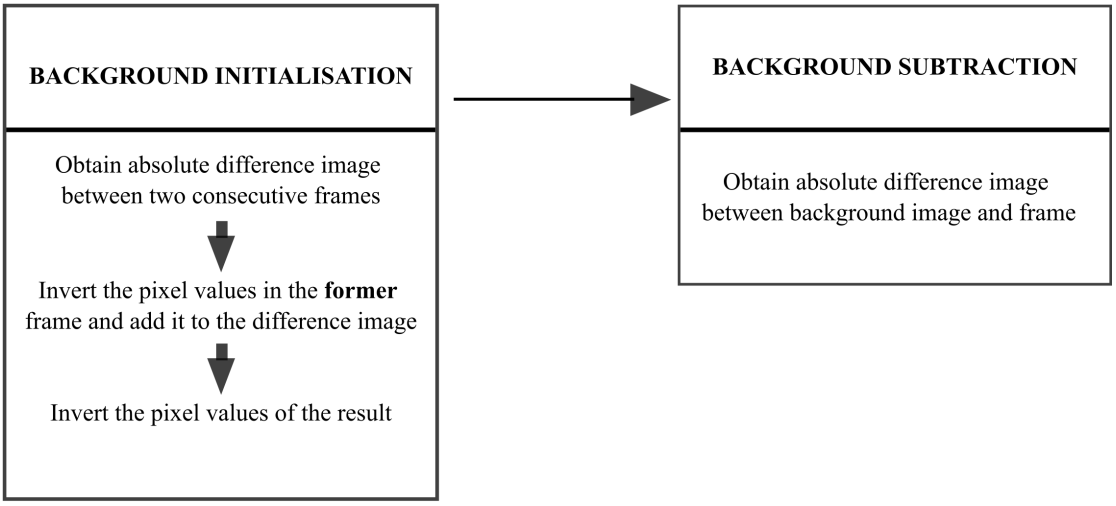


FIGURE 3.12: Overview of the proposed Background subtraction method

The Candidate Generation steps are summarised in Algorithm 2. Figs. 3.13 and 3.14 provide a visual on the flow of the algorithm.

Algorithm 2 Background Subtraction

1: **Input:** Sequence of U_k images $(M_g)_{0 \leq g \leq U_{k-1}}$,
2: **Output:** Sequence of U_k images $(\mathcal{LC}_g)_{0 \leq g \leq U_{k-1}}$
3: Obtain M_k and M_{k+step} from sequence
4: calculate D_k using Eq. 3.14
5: calculate background image $\mathcal{B}$ using Eq. 3.15
6: **for** $p \leftarrow 0$ **to** U_{k-1} **do**
7: $\quad \mathcal{LC}_p = |\mathcal{B} - M_p|$
8: **end for**

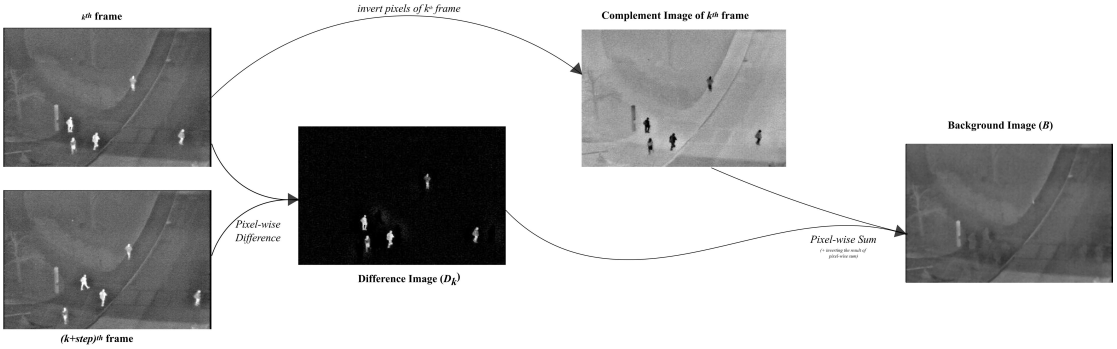


FIGURE 3.13: Generating a background image from two consecutive video frames

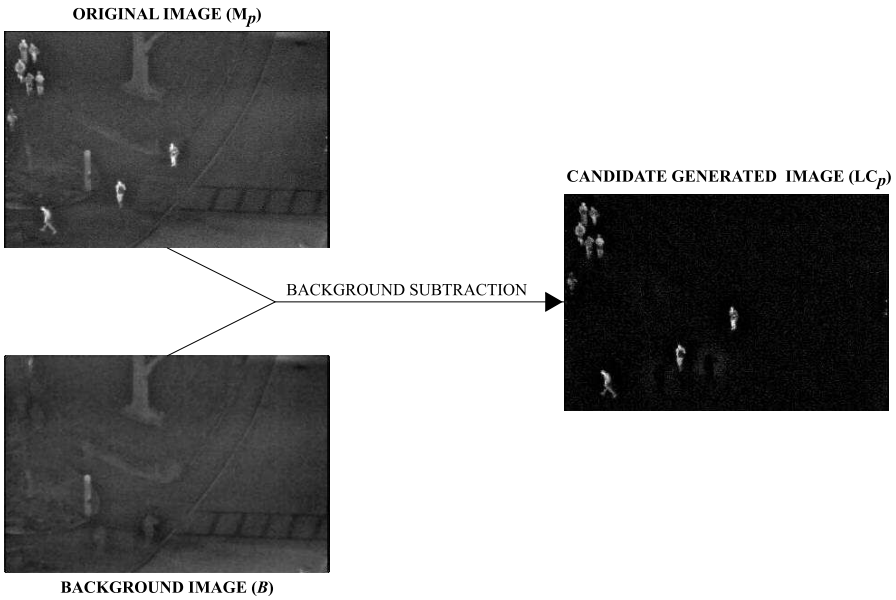


FIGURE 3.14: Candidate Generation after Background Subtraction

3.5 Motion-Constrained Graph Cut

Before describing the proposed method, an introduction to Graph Cut and the type of problem it solves is presented. Graph cut is from the family of optimisation methods called combinatorial optimization used for solving inference problems on discrete graphs formulated as finding the Maximum *à Posteriori* estimate (MAP) of a Markov Random Field (MRF) [78, 79]. Labelling problems are inference problems solved by combinatorial optimization. Greig et al. [80] were the first to produce an exact MAP estimate using min-cut/max-flow theorem in [81] to solve a binary labelling problem and were also the first to introduce Graph Cut to image processing by applying the exact solution to image restoration. Boykov and Jolly [24] were the first to extend the method in [80] to image segmentation.

Formally, the labelling problem is a function that maps observed data to labels. For our purposes, the observed data is the image and the labels are the classes. Optimization algorithms like Graph Cut perform efficient searches for the optimal labels among the possible set of labels. The optimal solution is determined by the energy function defining the labelling problem.

In this section, a motion-constrained Graph Cut energy function is presented for

pedestrian detection in IR images. The novelty of the proposed method is explained as follows. Let $\mathcal{Z}$ = ('ped', 'bkg') be labels assigned to a pixel corresponding to the ROI and background respectively. The labelling of $\mathcal{X}$ over $\mathcal{Z}$ is a function $h : \mathcal{X} \rightarrow \mathcal{Z}$. h_x specifies the label assignments to x in $\mathcal{X}$ and is taken from $\mathcal{Z}$. The formulation in [24] presents an energy function E incorporating a region $D(h)$ and boundary term $S(h)$ shown as follows

$$E(h) = \lambda \cdot \ D(h) \ + \ J(h) \tag{3.17}$$

where

$$D(h) = \sum_{x \in \mathcal{X}} D_x(h_x)$$

$$J(h) = \sum_{\{x,y\} \in \mathcal{N}} J_{x,y}(h_x, h_y) \cdot \epsilon(h_x, h_y)$$

and

$$\epsilon(a, b) = \begin{cases} 1, & \text{if } a \neq b \\ 0, & \text{otherwise} \end{cases}$$

where $\mathcal{N}$ are unordered pairs of neighbouring pixels from a standard neighbourhood system e.g., 4-, 8- or 26- neighbourhood system and λ is used to balance the contribution of the region and boundary term to the final segmentation result.

$D(h)$ measures how well pixels fit into the object or background models. The costs are calculated as

$$\begin{aligned} D_x(\text{`obj'}) &= -\ln \Pr(q_x|\text{`obj'}) \\ D_x(\text{`bkg'}) &= -\ln \Pr(q_x|\text{`bkg'}) \end{aligned} \tag{3.18}$$

where q_x is the intensity value of pixel x.

$J(h)$ is also called the smoothness term and was calculated as

$$J_{x,y}(h_x, h_y) = \exp\left(\frac{(q_x - q_y)^2}{2\sigma^2}\right) \tag{3.19}$$

.

$J_{x,y}(h_x, h_y)$ is higher when the pixels have similar intensity values as they will more likely be assigned the same label.

There are two areas where this formulation falls short in TIR images. Firstly, the low resolution and noisy nature of IR images mean that more importance will be given to the region term in many instances using this formulation. This means that a robust model for each class will have to be determined. From the literature, most models and approximated distributions do not generalise well across datasets, therefore, it is important to add another element to reduce over-dependence on the region term. Secondly, this formulation produces solutions where area with similar intensity values as the pedestrians are included in the solution irrespective of their location (shown in Fig. 3.15). The impact of the proposed energy is expressed in Fig. 3.16

A new energy function, referred to as a *motion-constrained Graph Cut energy (MCGCE)*, incorporates motion constraints and modifies Eq. 3.17. It is defined in Eq. (3.20) as

$$E(h) = D(h) + J(h) + M(h) \tag{3.20}$$

where

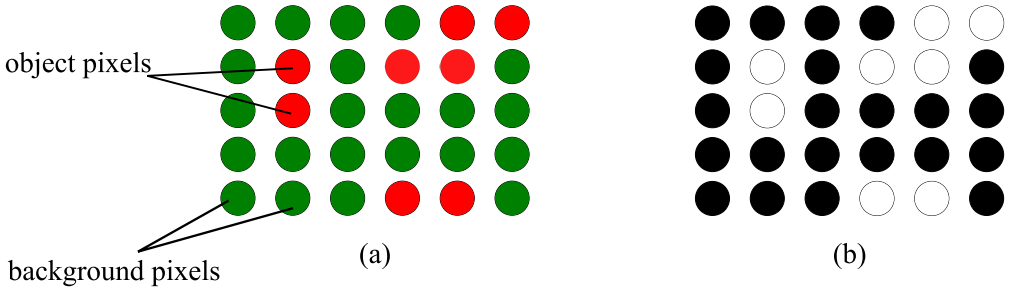


FIGURE 3.15: Topological Unconstrained solution:
The object pixels (shown as red in (a)) are properly labelled as foreground (shown as white in (b)) irrespective of their location

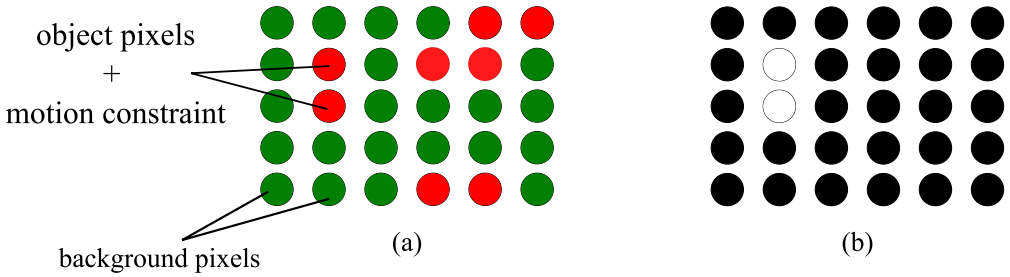


FIGURE 3.16: Constrained solution:
Only object pixels constrained by motion are labelled as foreground

$$
\begin{aligned}
D(h) &= \sum_{x \in \mathbf{X}} D_x(h_x) \ \cdot \delta(h_x) \\
J(h) &= \sum_{\{x,y\} \in \mathbf{N}} J_{x,y}(h_x, h_y) \cdot \epsilon(h_x, h_y) \\
M(h) &= \sum_{x \in \mathbf{X}} M_x(h_x) \cdot \beta(h_x)
\end{aligned}
$$

and

$$
\delta(h_x) = \begin{cases} 1, & \text{if } x \in \mathcal{T} \wedge x \notin \mathcal{D}comb \\ 0, & \text{otherwise} \end{cases}
$$

$$
\epsilon(h_x, h_y) = \begin{cases} 1, & \text{if } x \in \mathcal{T} \wedge y \in \mathcal{T} \wedge h_x \neq h_y \\ 0, & \text{otherwise} \end{cases}
$$

$$\beta(h_x) = \begin{cases} 1, & \text{if } x \in \mathcal{D}comb \\ 0, & \text{otherwise} \end{cases}$$

where $M(h)$ is the motion term, $\mathcal{T}$ is the set of pixels containing one or more motion pixels and $\mathcal{D}comb$ is the set of pixels with the highest energies from four directional difference images. To obtain $\mathcal{T}$, the image is divided into non-overlapping equal-sized detection windows such that only windows which have one or more pixels from $\mathcal{D}comb$ are considered by $D(h)$ and $S(h)$.

MCGCE was also inspired by the work in [23] where a framework was proposed which eliminates the need for separate modules for pedestrian detection and integrates appearance and motion patterns such that all the fine-tuning and adjustment happens during training. However, the framework of [24] is semi-supervised while that of [23] is supervised. Also, in the background subtraction method presented in Section 3.4, no information about the magnitude or direction of motion was included.

3.5.1 Motion Constraint

The motion constraint $\mathcal{D}comb$ provides an estimate of the location of each pedestrian in the image. Frame differencing detects moving regions and is commonly used in tracking algorithms [82]. The presence of motion can be obtained from the absolute difference between image pairs $\mathcal{D}_f$. The direction of motion can be obtained from the absolute difference $\mathcal{D}U_f$, $\mathcal{D}D_f$, $\mathcal{D}L_f$ and $\mathcal{D}R_f$ between the first image and shifted versions of the second image.

$$\mathcal{D}_f = |I_f - I_{f+1}|$$
$$\mathcal{D}U_f = |I_f - I_{f+1} \uparrow|$$
$$\mathcal{D}D_f = |I_f - I_{f+1} \downarrow|$$
$$\mathcal{D}L_f = |I_f - I_{f+1} \leftarrow|$$
$$\mathcal{D}R_f = |I_f - I_{f+1} \rightarrow|$$

During experiments, it was found that the energy of the image was highest when the image was shifted in the direction of motion and the least when shifted in the opposite direction. Also, because the surveillance footage is taken from different angles and there are usually several pedestrians are going in different directions, the energy for each subject is higher in at least two directions, that is, either in the $\uparrow$ or $\downarrow$ direction and either in $\leftarrow$ or $\rightarrow$ direction. Therefore, a new difference image $\mathcal{D}comb$ was created by combining the pixels with the highest energies from each directional difference image and is defined as

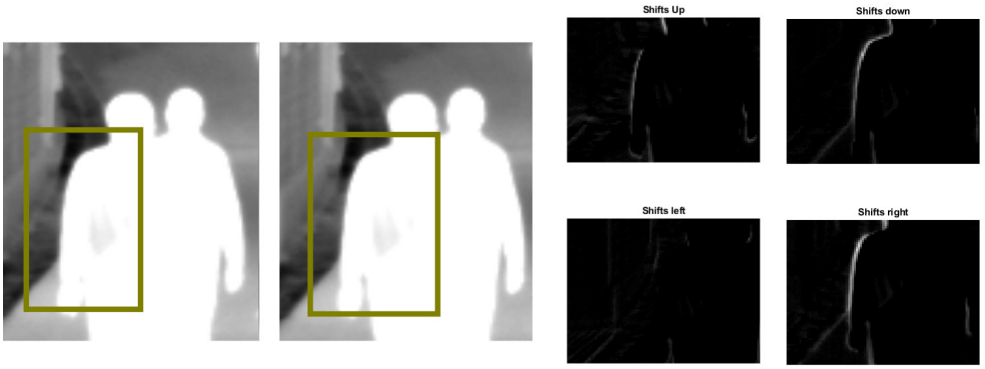
$$\mathcal{D}comb = \{(\mathcal{D}U_f > e) \cup (\mathcal{D}D_f > e) \cup (\mathcal{D}L_f > e) \cup (\mathcal{D}R_f > e)\} \tag{3.21}$$

where e is used to extract the highest energies from each directional difference image.

In Fig. 3.17, it can be seen how the shifted difference images provide information about the direction of motion. These findings are different from [23] where the image energy is least in the direction of motion. The hypothesis is that this might be caused by the difference in visible and thermal infrared images. Fig. 3.18 shows $\mathcal{D}comb$ and how it provides information about the location of the pedestrians.

3.5.1.1 Definition of M(h)

The term $M(h)$ in Eq. (3.20), the cost of assigning a label to a pixel determined by the motion constraint, is defined as follows:



(a) (b) (c)

FIGURE 3.17: (a) and (b) are two consecutive frames with an area of interest selected and (c) shows the directional difference images around that selected area. The image energy is higher when the image is shifted to the right than to the left, and then when it is shifted downwards than upwards. So, without previous knowledge, one can tell the pedestrian is moving to the right and slightly downwards.

$$M_x(\text{“ped”}) = \begin{cases} \eta, & \text{if } x \in \mathcal{D}comb \\ 0, & \text{otherwise} \end{cases} \tag{3.22}$$

$$M_x(\text{“bkg”}) = \begin{cases} \eta, & \text{if } x \notin \mathcal{D}comb \\ 0, & \text{otherwise} \end{cases} \tag{3.23}$$

The value of η can be any large number. This is to ensure that all pixels affected by the motion constraint are included for label assignment.

3.5.2 Graph Construction

An example of a graph constructed over an image is shown in Fig. 3.19. Let $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ be a graph with $\mathcal{V}$ as vertices and $\mathcal{E}$ as edges. The vertices represent the image pixels and the edges connect pixels to one another. Source S and sink T are two additional vertices that connect to pixels belonging to the object of interest and background respectively. The connection between pixels is determined by the neighbourhood system. In Fig. 3.19, a 4-neighbourhood system is used. The

(a)

(b)

(c)

(d)

(e)

(f)

FIGURE 3.18: (a) Image (b) $\mathcal{D}U_f$ (c) $\mathcal{D}R_f$ (d) $\mathcal{D}L_f$ (e)$\mathcal{D}D_f$ (f) $\mathcal{D}comb$

assignment of weights to edges is determined by the terms of the energy function. Table 3.1 gives the edge weights for the graph.

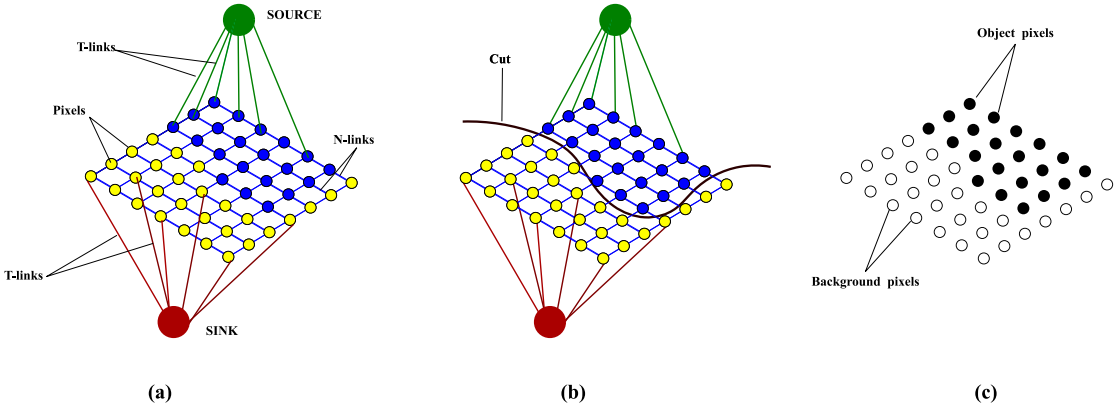


FIGURE 3.19: Binary labelling of an image using Graph Cut (a) shows the graph constructed from the image (b) shows the minimum cut separating the vertices (c) shows the binary labelling as a result of the cut

TABLE 3.1: Edge weights of the graph constructed from the image

Edge	**Weight**	**for**
$\{x, y\}$	$J_{x,y}$	$\{x, y\} \in \mathcal{N}, x \in \mathcal{T} \wedge y \in \mathcal{T}$
$\{x, S\}$	$D_x(\text{“ped”})$	$x \in \mathcal{X}, x \in \mathcal{T}$
	$M_x(\text{“ped”})$	$x \in \mathcal{X}, x \in \mathcal{D}comb$
	0	$x \in \mathcal{X}, x \notin \mathcal{D}comb \wedge x \notin \mathcal{T}$
$\{x, T\}$	$D_x(\text{“bkg”})$	$x \in \mathcal{X}, x \notin \mathcal{T}$
	$M_x(\text{“bkg”})$	$x \in \mathcal{X}, x \notin \mathcal{D}comb$
	0	$x \in \mathcal{X}, x \notin \mathcal{D}comb \wedge x \notin \mathcal{T}$

3.5.3 Energy Minimization

Following the energy formulation, the graph is constructed, weights are assigned and the minimum cut separating the two terminals is computed which produces the final labelling on the original image (see Fig. 3.19c). The energy is minimized using the algorithm proposed by [83].

3.6 Summary

Details of the dataset selected and the Candidate Generation and Validation techniques formulated in this research were presented in this chapter.

The dataset chosen for evaluation ranges from the earliest and most extensively used for pedestrian detection to modern databases. LTIR is the most modern database and has higher resolution images while OSU is the earliest and has lower resolution and more noisy images. They were also chosen to reflect a wide range of scenarios, weather conditions, pedestrian sizes and camera viewpoints.

In the entropy-based histogram specification algorithm, the principle of minimum cross-entropy was extended to include principles of Histogram specification by using energy loss to modify an image's histogram distribution. The success of the method lies in the fact that Entropy is more when information is uniformly distributed. By increasing the entropy within the human regions, the areas of lower entropy are iteratively eliminated until the desired goal is attained. Histogram equalisation is responsible for increasing the entropy within the human regions while minimum cross-entropy provides the criteria for subsequent histogram modification until convergence is reached.

In the Background Subtraction method, a computationally inexpensive 2-frame background initialisation method was put forward. The success of the method lies in the fact that infrared images have large areas of similar pixel intensities and do not experience sudden appearance changes like visible camera footage. The low resolution and noisy nature of infrared images mean that only slight modifications in the pixel values where motion is detected are necessary to obtain a background able to detect pedestrians in other video frames.

The Candidate Validation technique aims to eliminate the separate modules for candidate generation and validation. The semi-supervised nature of the algorithm is similar to that in [24] where pixels are selected to build the appearance model of the image, however, motion constraint is incorporated to guide the final segmentation result.

In the next chapter, detailed experimental results and discussions will be presented on each method put forward in this chapter.

Chapter 4

Experimental Results and Discussion

4.1 Introduction

This chapter presents the experimental setup, results and discussions of the proposed candidate generation and validation techniques put forward in this research for pedestrian detection in IR images.

4.2 Experimental Setup

4.2.1 Framework Development Environment

The algorithms for candidate generation and validation were implemented using MATLAB R2018a^{TM} on an Intel i7-4790 CPU 3.60GHz with 8GB RAM.

4.2.2 Performance Evaluation Measures

The performance measures are Total True Positive (TTP), Total False Positive (TFP), Positive Predictive Value (PPV) and Sensitivity.

A detection is said to be False Positive (FP) when a non-pedestrian is detected as a pedestrian and True Positive (TP) when a pedestrian is correctly detected. Therefore, TFP is the sum of regions incorrectly included in the detection results and TTP is the sum of correctly detected pedestrians. Sensitivity is the algorithm's ability to detect pedestrians. A high sensitivity indicates high performance. PPV measures the rate of detecting false positives. A high PPV value is desired because that translates to a low rate of false positives detected by the algorithm. Some methods in the literature use precision and recall but the values obtained correspond to the PPV and sensitivity respectively, therefore, where required, they are used interchangeably. Sensitivity and PPV are computed as

$$\text{Sensitivity} = \frac{\#\text{TTP}}{\#\text{TNP}} * 100 \tag{4.1}$$

$$\text{PPV} = \left(1 - \frac{\#\text{TFP}}{\#\text{TNP}}\right) * 100 \tag{4.2}$$

4.3 Experimental Results and Discussions

4.3.1 Qualitative Evaluation

The qualitative performances are presented to showcase the improvements made by the techniques proposed in this study. The Entropy-based histogram modification (EHM) is compared with the Minimum Cross-Entropy algorithm in [36]. The motion-constrained Graph Cut framework is compared with the framework in [24]. For the Background Subtraction Method (BS2FI), the results show the two frames

used for initialisation, the background image, the image to be subtracted and the candidate image generated. Each method uses images from the four databases detailed in Section 3.2.

4.3.1.1 Entropy-based histogram modification

The Entropy-based histogram modification algorithm is visually compared to the Minimum Cross Entropy algorithm (MCE) [36] on four different databases. Fig. 4.1 and 4.2 present the output on images from the LTIR database which has six sequences for pedestrian detection. Fig. 4.3 and 4.4 presents the output on one image from each of the nine sequences in the LITIV database. Fig. 4.5 and 4.6 present the result on one image from each of the ten sequences of the OSU thermal database. Fig. 4.7 presents the results from the TMIR database.

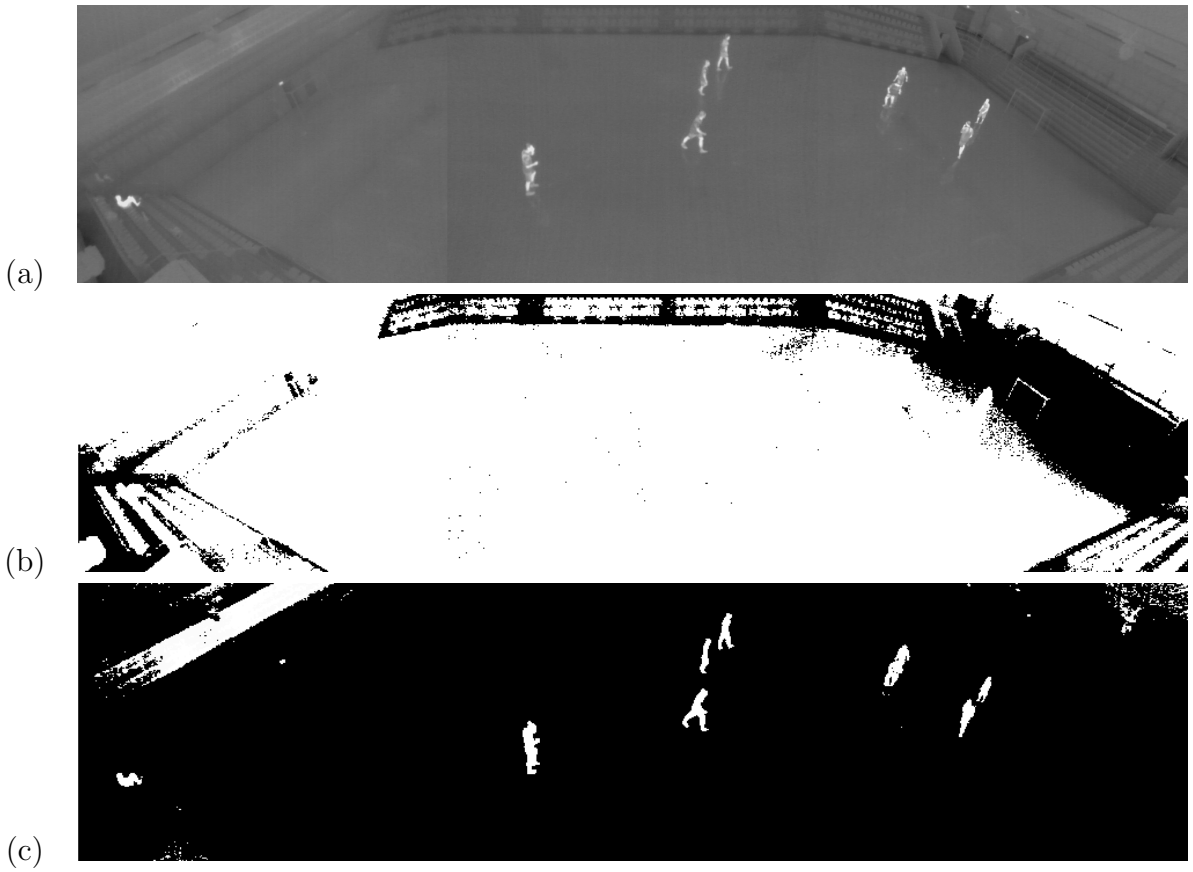


FIGURE 4.1: EHM results on the LTIR database (a) Image (b) MCE (c) EHM

The threshold chosen for the images from the LTIR database are not sufficient to separate the pedestrians from the background even though the pedestrians

(a) (b) (c)

FIGURE 4.2: EHM results on the LTIR database (a) Image ((b) MCE (c) EHM

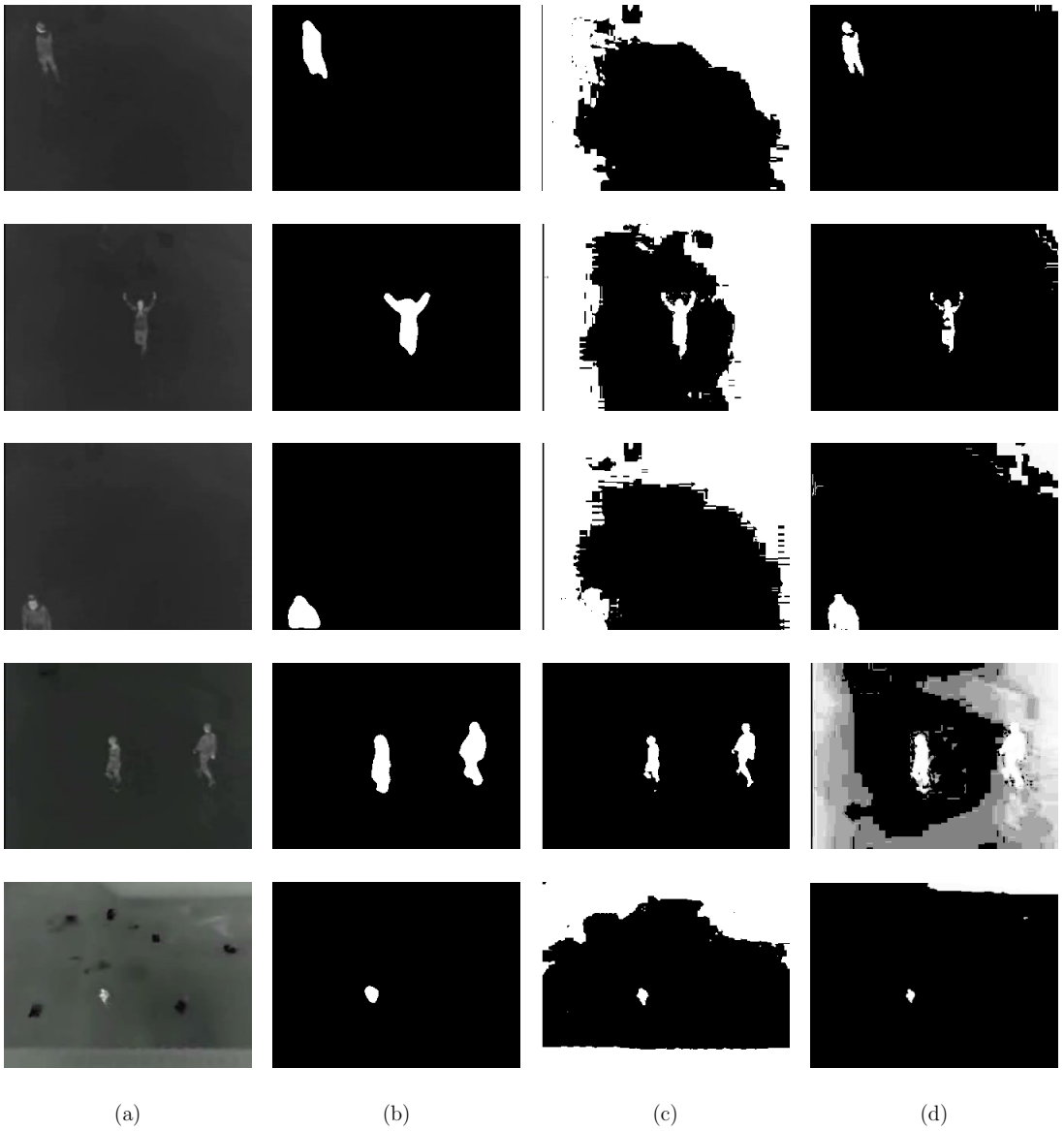
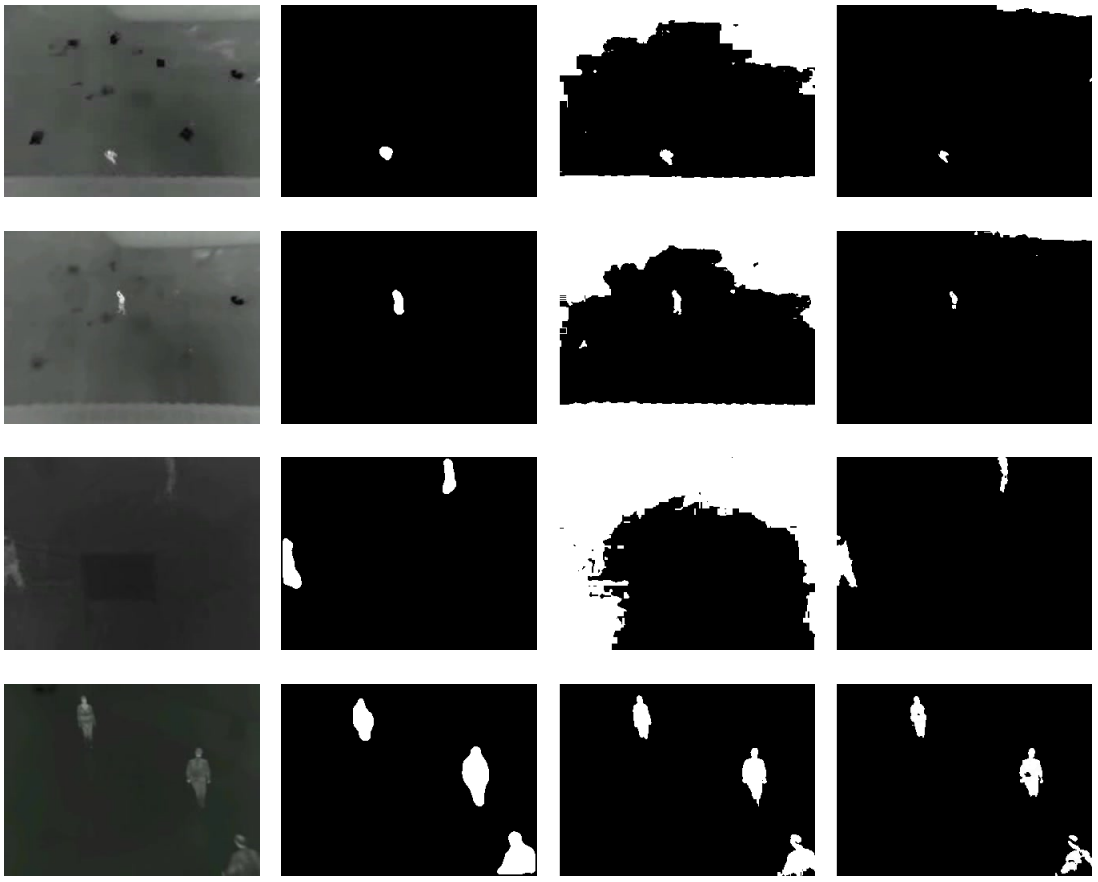


FIGURE 4.3: EHM results on the LITIV database (a) Image (b) Ground-truth (c) MCE (d) EHM

are sufficiently brighter than the background. However, the proposed method performs poorly on the Mixed Distractors Sequence (Row 5 in Fig. 4.2) as the infrared emissions vary widely on the pedestrians.

In the case of the LITIV dataset, even though the background appears uniformly black to the eye in most images, this is not the case. In Fig. 3.8, the histogram of images from the LITIV database are the narrowest which means that a lot of information is compressed within a very narrow range of pixel intensities. Thus, the background of the images creates a problem for minimum cross-entropy and

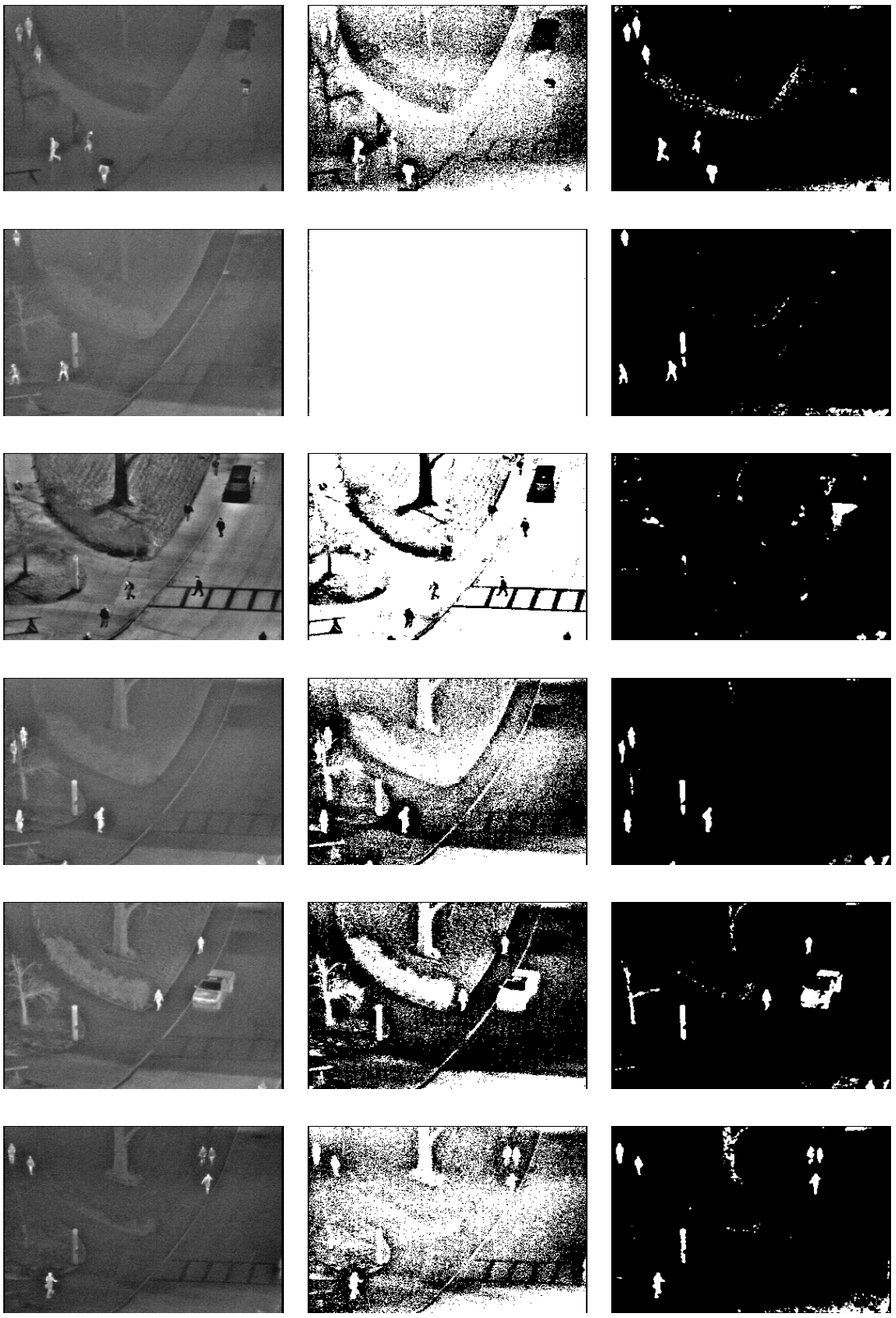


(a) (b) (c) (d)

FIGURE 4.4: EHM results on the LITIV database (a) Image (b) Ground-truth (c) MCE (d) EHM

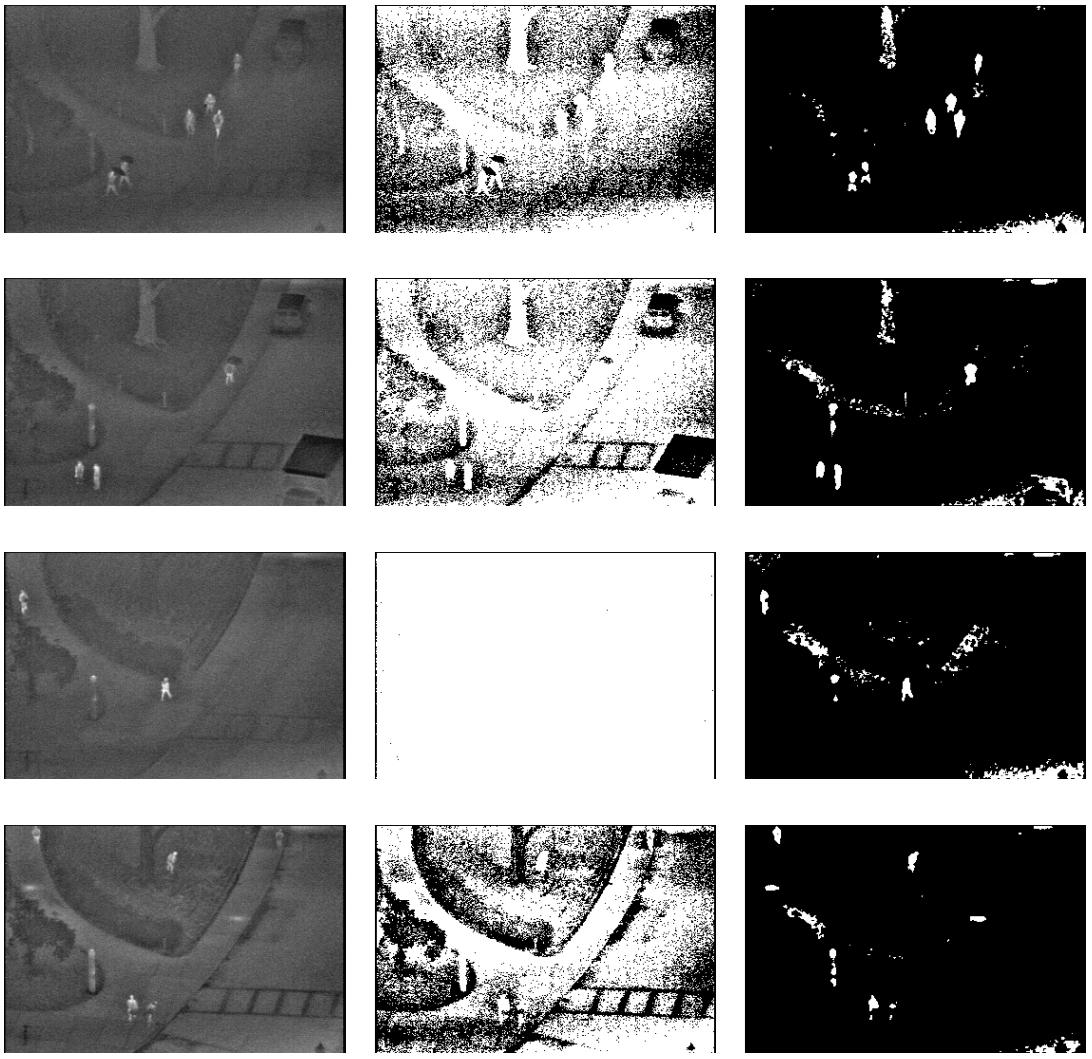
pedestrians at the edge of the images cannot be extracted. The method performs poorly on Sequence 4 (row 4 in Fig, 4.3) because the algorithm converges after one iteration and the initial value is too low to properly separate the pedestrian from the background. In the case of this image, both algorithms chose the same threshold value of 61, but the difference in output is due to histogram equalisation.

Unlike the LTIR and LITIV databases, it can be seen that there is a more varied response to the use of minimum cross-entropy in the OSU thermal dataset. In Sequences 2 and 9 (row 2 in Fig. 4.5 and row 3 in Fig. 4.6), the threshold selected by minimum cross-entropy was so low that the whole image was selected. The only sequence that the proposed method performed poorly on is row 3 in Fig. 4.5 where the pedestrians are not bright.



(a) (b) (c)

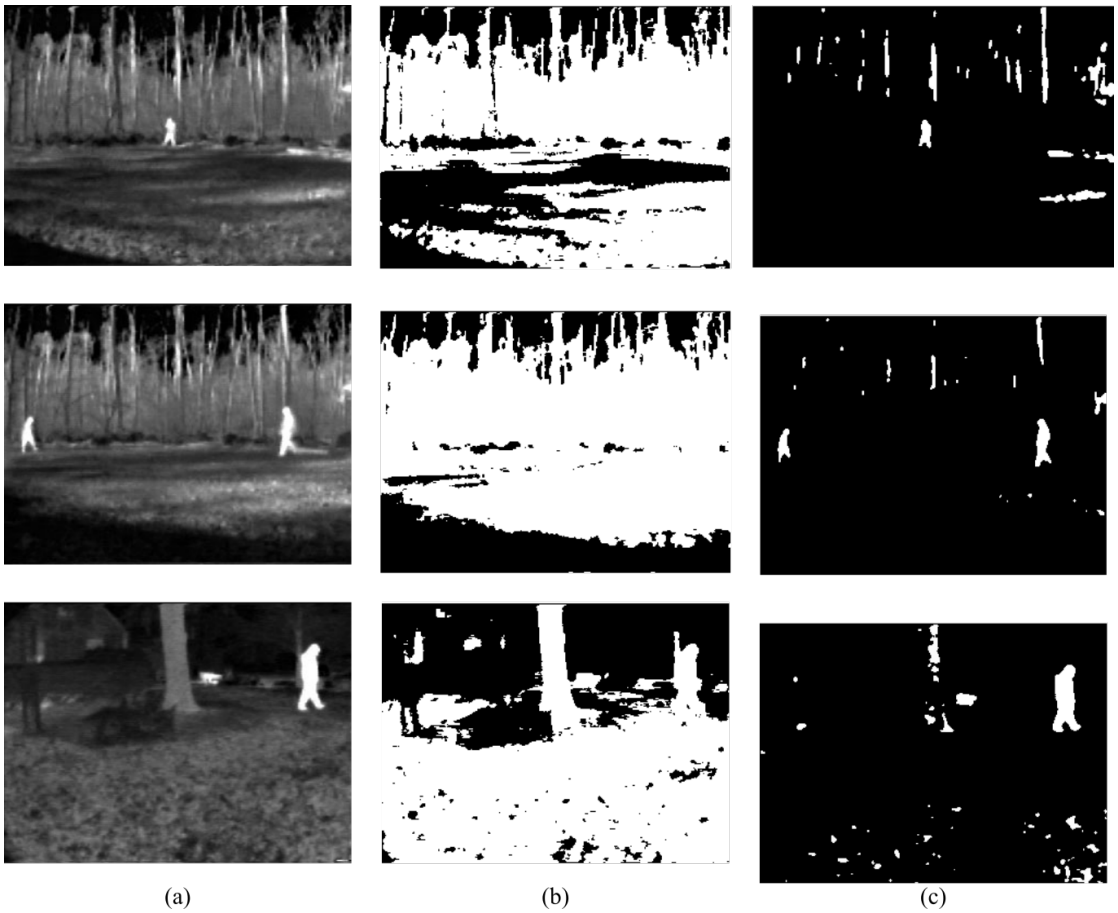
FIGURE 4.5: EHM results on the OTCBVS (OSU) Thermal database (a) Image (b) MCE (c) EHM



(a) (b) (c)

FIGURE 4.6: EHM results on the OSU Thermal database (a) Image (b) MCE (c) EHM

In the TMIR database, it can be seen that the pedestrian is lost within the vegetation when Minimum Cross-Entropy thresholding is used while the proposed method can extract the pedestrians.



(a) (b) (c)

FIGURE 4.7: EHM results on the TMIR database (a) Image (b) MCE (c) EHM

4.3.1.2 Background Subtraction using 2-frame Initialisation

Fig. 4.8, 4.9, 4.10 and 4.11 show the results obtained using the proposed Background Subtraction algorithm (BS2FI). Each figure has five rows showing the two frames M_k and $M_{(k+step)}$ selected to obtain the background image, the background image $\mathcal{B}$ obtained after the pixel-wise difference and sum operations, the image M_p from which pedestrians are to be extracted and the image $\mathcal{LC}_p$ showing the candidates generated. The columns are different images chosen from the database under consideration. The *step* in $M_{(k+step)}$ was determined empirically from the rate of motion and size of pedestrians. The method performs well on all four databases as it can extract the pedestrians from the background and all the pedestrians in the original image are accounted for in the candidate generation image.

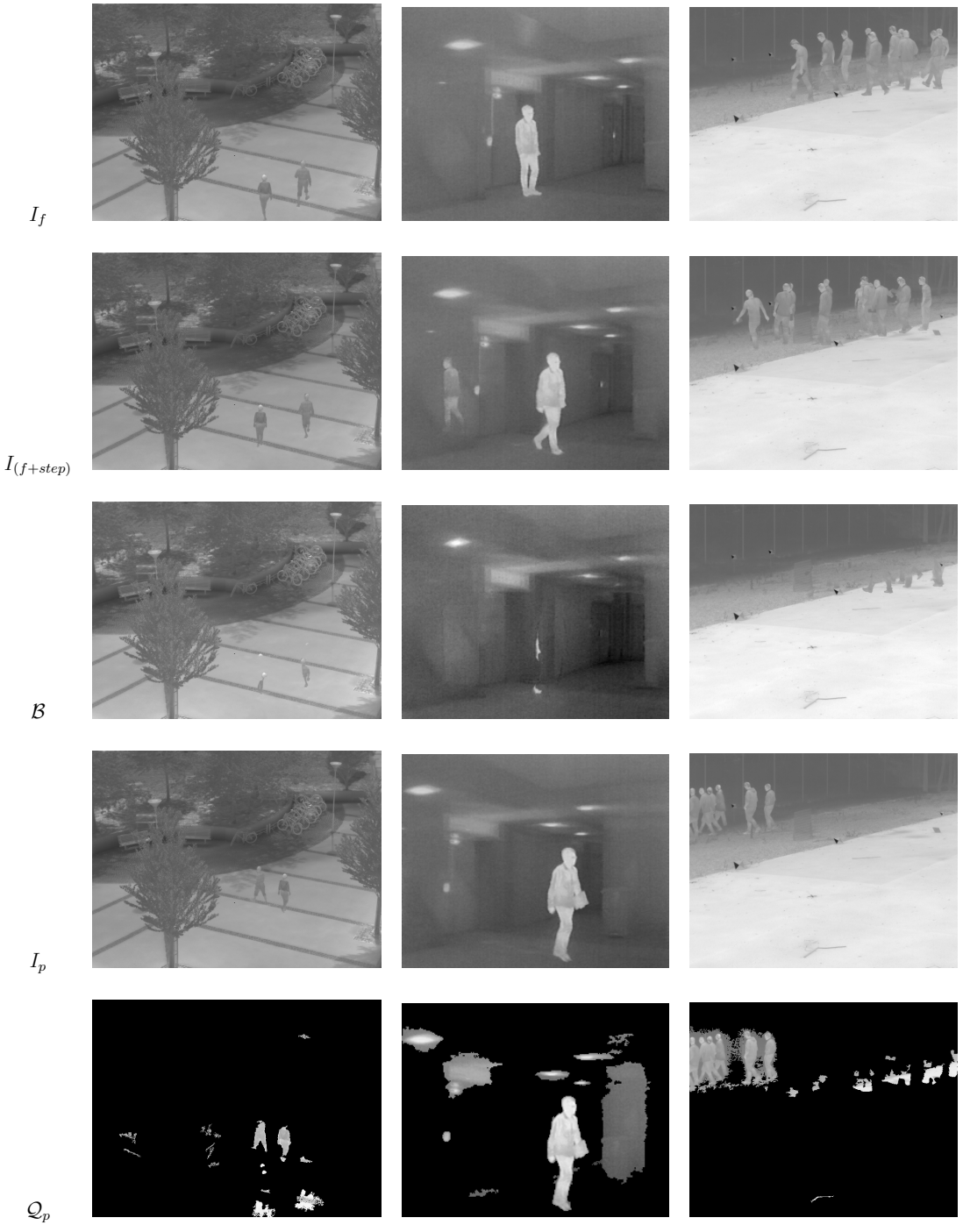


FIGURE 4.8: BS2FI results on the LTIR database. Columns (a), (b) and (c) are different images chosen from the database. The third row is the background image obtained using the images in the first two rows. The fifth row contains the pedestrian candidates to be forwarded for validation obtained from the difference between images in third and fourth rows.

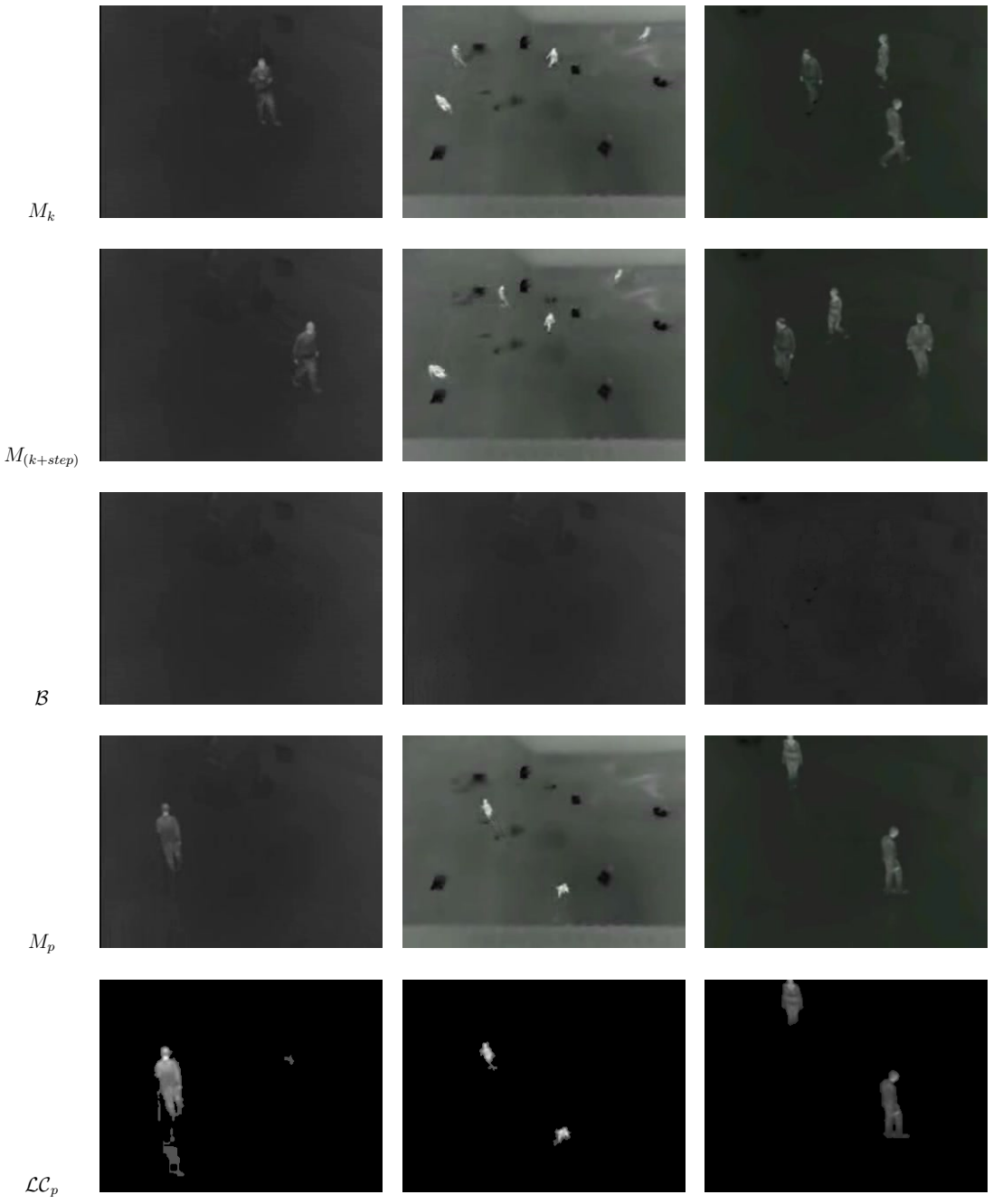


FIGURE 4.9: BS2FI results on the LITIV database. Columns (a), (b) and (c) are different images chosen from the database. The third row is the background image obtained using the images in the first two rows. The fifth row contains the pedestrian candidates to be forwarded for validation obtained from the difference between images in third and fourth rows.

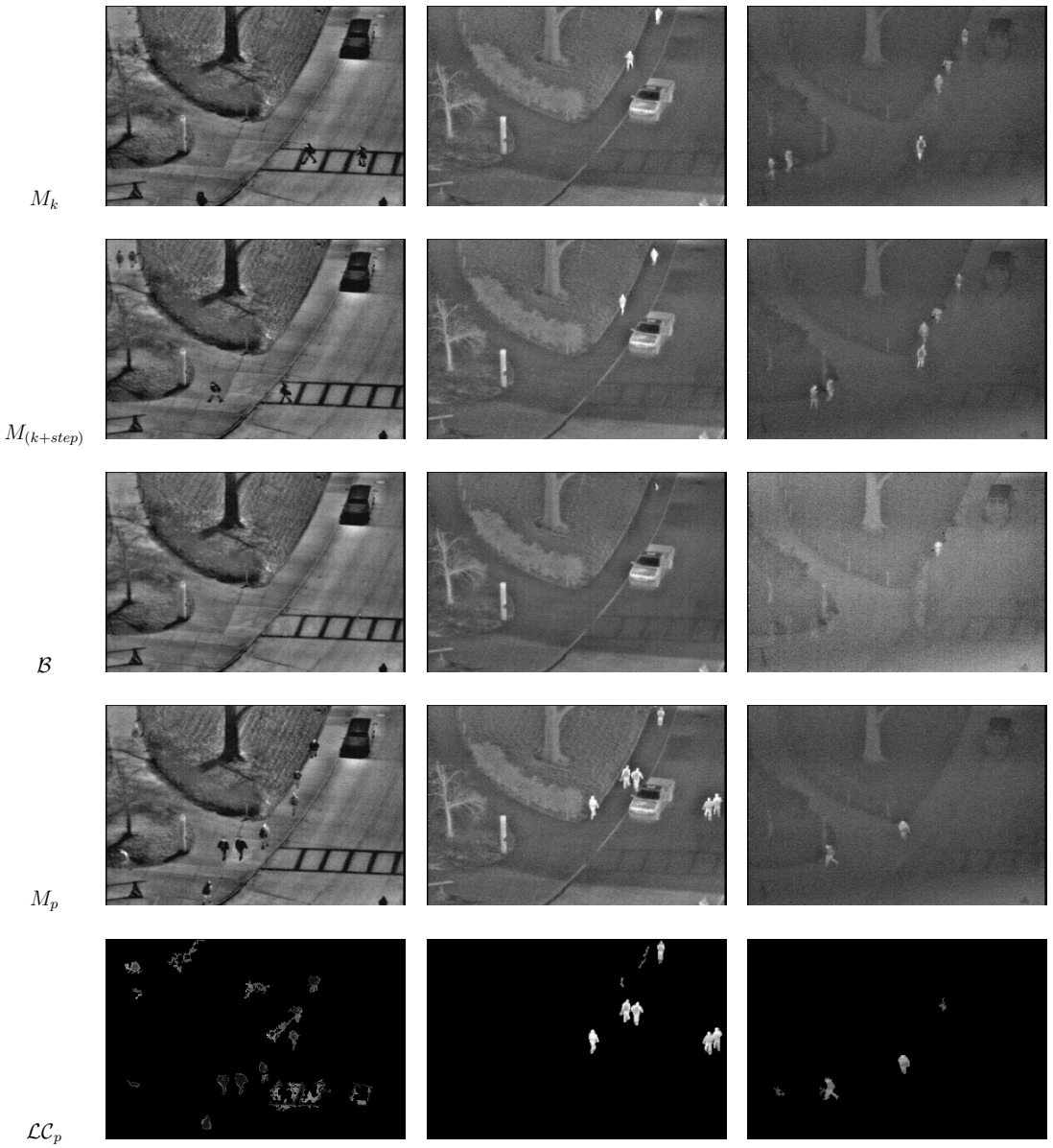


FIGURE 4.10: BS2FI results on the OSU database. Columns (a), (b) and (c) are different images chosen from the database. The third row is the background image obtained using the images in the first two rows. The fifth row contains the pedestrian candidates to be forwarded for validation obtained from the difference between images in third and fourth rows.

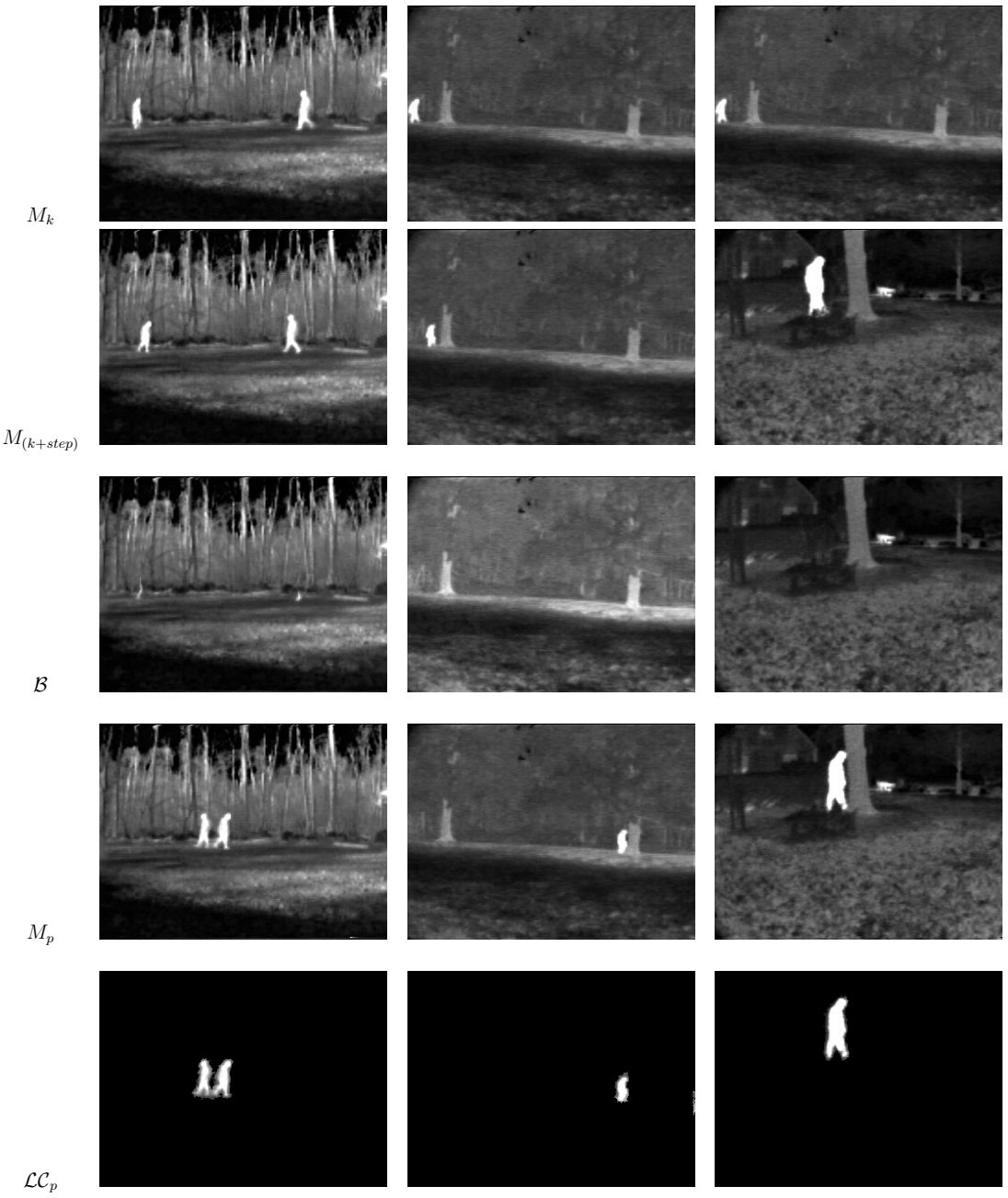
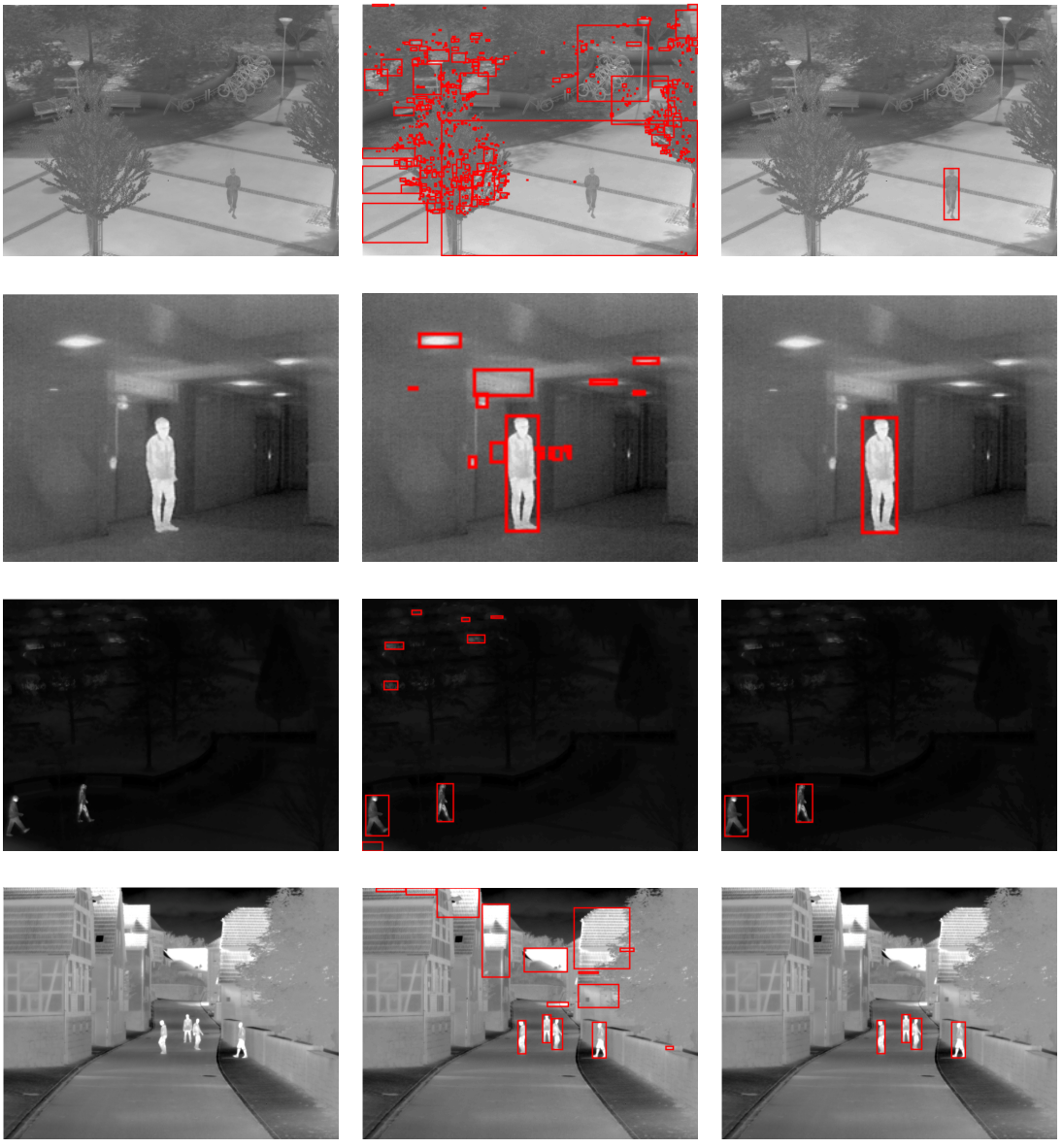


FIGURE 4.11: BS2FI results on the TMIR database. Columns (a), (b) and (c) are different images chosen from the database. The third row is the background image obtained using the images in the first two rows. The fifth row contains the pedestrian candidates to be forwarded for validation obtained from the difference between images in third and fourth rows.

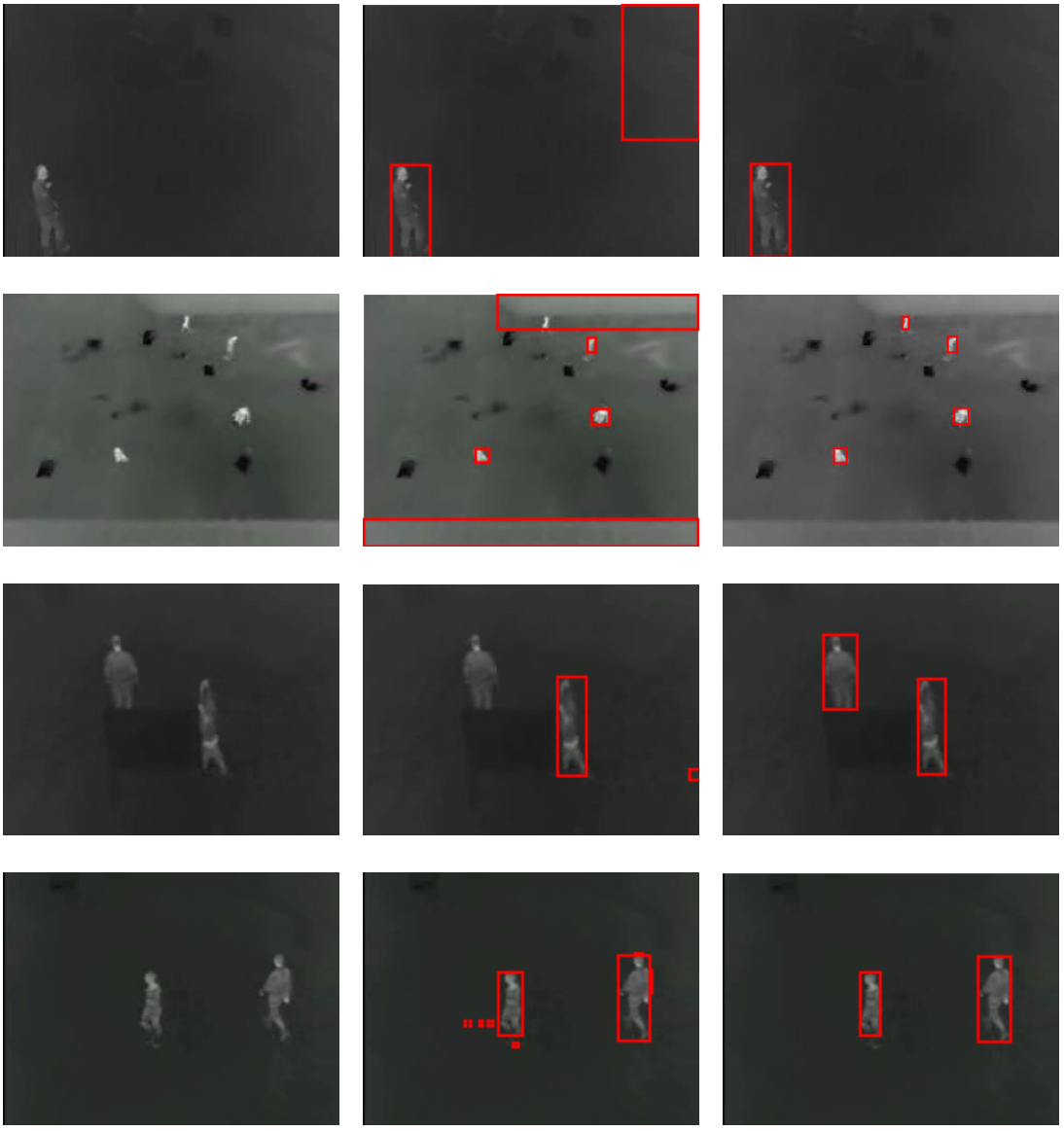
4.3.1.3 Motion-constrained Graph Cut

The Motion-constrained Graph Cut framework (MCGCE) is visually compared with the method in [24] (GC). The results are shown in Figs. 4.12, 4.13, 4.14 and 4.15. In the LTIR dataset, MCGCE improves the detection results by eliminating objects with similar intensities. In the first row of Fig. 4.12, the lines patterns on the road are similar in intensity to the pedestrian Therefore, the GC algorithm labels the pedestrian and all the lines connected to the pedestrian as one object. However, with the motion constraint, the pedestrian is detected using the MCGCE algorithm. The LITIV database also has such problems when the pedestrian is connected to the object with similar intensity when using GC. Although in most cases it detects the pedestrians, GC detects erroneous regions as well. In the OSU database, many FP regions get included in the detection results because of noise when using the GC algorithm. However, we see a drawback of MCGCE where the car is included in the results. In the TMIR database, the pedestrians are uniformly bright and present little challenge to being detected by both GC and MCGCE, but in the GC results, the background objects with similar intensity get included.



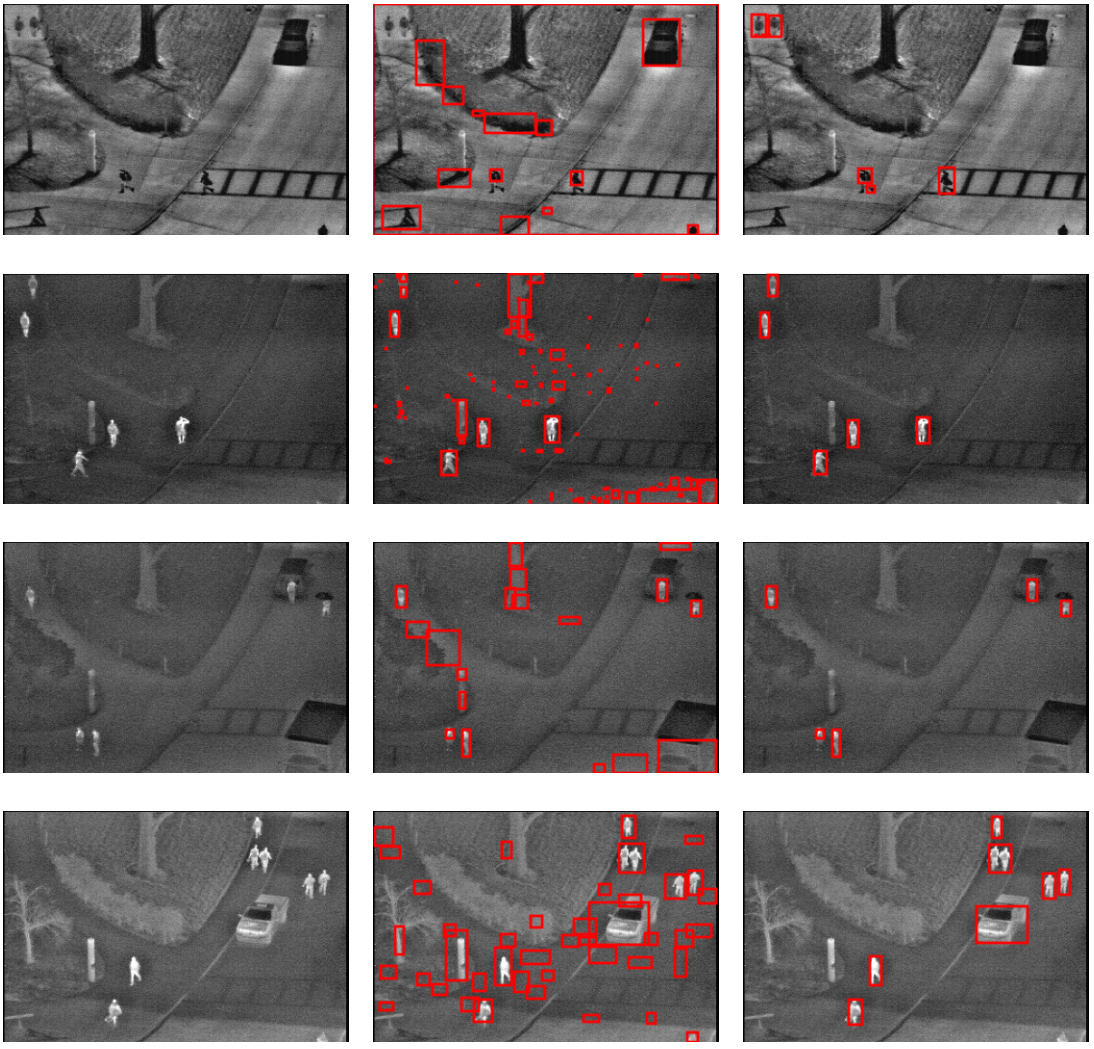
(a) (b) (c)

FIGURE 4.12: MCGCE results on LTIR database (a) Image (b) GC (c) MCGCE



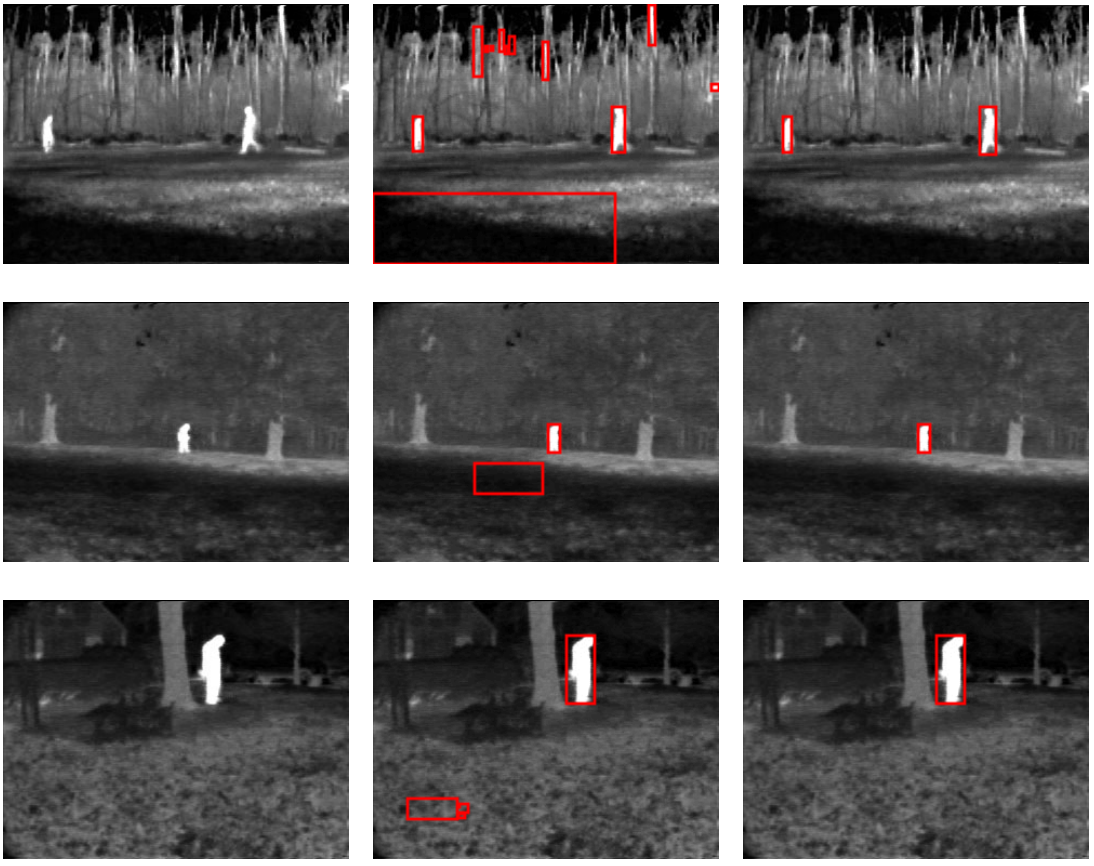
(a) (b) (c)

FIGURE 4.13: MCGCE results on LITIV database (a) Image (b) GC (c) MCGCE



(a) (b) (c)

FIGURE 4.14: MCGCE results on OSU database (a) Image (b) GC (c) MCGCE



(a) (b) (c)

FIGURE 4.15: MCGCE results on TMIR database (a) Image (b) GC (c) MCGCE

4.3.2 Quantitative Evaluation

The performance measures described in section 4.2.2 will be used for the quantitative evaluation. For EHM, sequences are chosen from the LTIR and OSU dataset where pedestrians are brighter than the background. For BS2FI and MCGCE, sequences are chosen from the LTIR dataset where pedestrians are in constant motion most of the time regardless of the polarity of the pedestrian. Tables 4.1 and 4.2 show the ability of the Entropy-based Histogram Modification (EHM) and Background Subtraction (BS2FI) methods to correctly detect pedestrians calculated using sensitivity. Table 4.3 presents the sensitivity and PPV of the motion-constrained Graph Cut (MCGCE).

TABLE 4.1: Sensitivity Results for the Entropy-based Histogram Modification algorithm

Database	**Sequence**	**Sensitivity**
OSU	1,2,4-10	98.14
TMIR	11 (OMT)	99.30
LITIV	All	98.91
LTIR	Saturated	96.44
LTIR	Street	98.78
LTIR	Crossing	99.95
LTIR	Hiding	99.63
LTIR	Mixed Distractors	95.68
LTIR	Soccer	**99.97**

TABLE 4.2: Sensitivity Results for the BS2FI algorithm

Database	**Sequence**	**Sensitivity**
OSU	All	94.69
TMIR	11 (OMT)	**100.00**
LITIV	All	**100.00**
LTIR	Street	99.52
LTIR	Crossing	98.73
LTIR	Quadrocopter2	98.97
LTIR	Soccer	98.59
LTIR	depthwise crossing	99.35

As the OSU dataset has been extensively tested in the literature over the years and has the most comprehensive ground-truth, the results of the proposed algorithms

TABLE 4.3: Sensitivity and PPV Results for the Motion-Constrained Graph Cut (MCGCE) algorithm

Database	**Sequence**	**Sensitivity**	**PPV**
OSU	All	97.65	99.65
TMIR	11 (OMT)	**100.00**	**100.00**
LITIV	All	99.59	99.75
LTIR	Street	98.78	99.83
LTIR	Crossing	98.93	100.00
LTIR	Quadrocopter2	99.97	99.55
LTIR	Soccer	99.59	100.00
LTIR	depthwise crossing	99.35	99.79

are presented in comparison with other methods that make use of it. These results are shown in Tables 4.4, 4.5 and 4.6. The Candidate generation methods are paired with validators to facilitate a comparison with other methods in the literature. EHM is paired with the unsupervised validation method in [61]. BS2FI is paired with the framework in [24] as put forward in [84]. MCGCE is a semi-supervised single model for pedestrian detection formulated to eliminate the need for separate modules for candidate generation and validation. The third sequence of the OSU database is omitted to accommodate the EHM algorithm.

TABLE 4.4: Comparing BS2FI, EHM and MCGCE with other methods using Total True Positives (TTP) on the OSU dataset

Sequence	#Pedestrians	[16]	[28]	[65]	[85]	[86]	BS2FI	EHM	MCGCE
1	(91)	88	**90**	87	77	78	85	88	**90**
2	(100)	94	95	96	99	98	97	**100**	98
4	(109)	107	108	**109**	107	**109**	**109**	**109**	98
5	(101)	90	95	100	97	**101**	97	**101**	109
6	(97)	93	94	94	92	**97**	93	94	99
7	(94)	92	93	86	78	80	90	93	**94**
8	(99)	75	80	97	89	96	93	**98**	92
9	(95)	**95**	**95**	**95**	91	**95**	**95**	**95**	**95**
10	(97)	**95**	**95**	94	91	83	89	89	**95**
1-10	(883)	829	845	858	821	829	848	**867**	863

TABLE 4.5: Comparing BS2FI, EHM and MCGCE with other methods using Total False Positives (TFP)

Sequence	[16]	[28]	[65]	[85]	[86]	BS2FI	EHM	MCGCE
1	0	0	5	3	0	0	0	0
2	0	0	14	2	2	2	0	1
4	1	0	18	7	10	0	0	0
5	0	0	13	16	16	1	2	0
6	0	0	2	8	0	0	0	0
7	0	0	4	8	0	1	1	0
8	1	1	3	8	0	0	0	0
9	0	0	2	4	0	0	0	1
10	3	3	8	18	16	0	0	0
1-10	5	4	69	74	44	4	3	**2**

TABLE 4.6: Comparison of Precision and Recall of the proposed methods with the state-of-the-art on the OSU dataset

Method	**Precision**	**Recall**
[20]	0.8600	0.8900
[69]	0.7100	0.6100
[87]	0.9920	0.9775
BS2FI	0.9922	0.9469
MCGCE	0.9965	0.9765
EHM	0.9967	0.9814

4.4 Summary

The performances of the candidate generation and validation techniques formulated in this study have been tested on four publicly available databases and the results have been presented in this chapter. The results show that the adaptations made to the original visible image algorithms in this study make them produce better results when used on thermal images for pedestrian detection. The results also show that their performance is comparable to methods in the literature.

In the next chapter, the conclusions and recommendations for future work will be presented.

Chapter 5

Conclusion and Recommendations for Future Work

5.1 Summary and Contribution

In this research, a detailed review of the literature was carried out on candidate generation and validation techniques used for pedestrian detection in infrared images. Many of the existing methods in the literature make use of algorithms created for visible light images, but due to the differences in imaging characteristics between visible and thermal infrared images, the algorithms do not perform well. In addition, many of the methods that attempt to create algorithms specific to infrared images only perform well on one or two databases usually because the algorithm has approximated or assumed a distribution that suits the data under observation. Attempting to create algorithms that generalise across, supervised techniques are increasingly being adopted for pedestrian detection in the thermal domain. However, these methods still use features or pre-trained models developed for visible images, and the original problem of differences in imaging characteristics persists. Some authors have

equally put forward models trained using only infrared images, and the results have been that performance is higher when the test data is similar to the training data which is similar to the unsupervised technique of approximating or assuming a distribution. Furthermore, the disparity in infrared images from different databases comes not only from the disparity in the scene to be captured such as weather conditions but from the camera used for capture and the settings used when capturing. Finally, methods in the literature for pedestrian detection fall into two groups, unsupervised and supervised. However, the review shows that there is still a significant dependence on the contrast between the pedestrian and the background especially when saliency is used to build attention mechanisms in either approach.

The motivation of this research was to incorporate strategies into existing frameworks used in visible image processing techniques for pedestrian detection without the need to initially assume a model for the image distribution. Three algorithms were put forward, two for candidate generation and one for candidate validation. The candidate generation techniques formulated include an Entropy-based histogram modification algorithm and a Background subtraction method featuring a 2-frame initialisation method. The candidate validation technique makes use of a motion-constrained Graph Cut energy function.

The Entropy-based histogram modification algorithm incorporates a strategy for energy loss to iteratively modify the histogram of an image for background elimination and pedestrian retention. This algorithm was motivated by the fact that the dynamic range of an infrared image controls how much information is displayed and was developed by combining histogram equalisation with minimum cross-entropy for minimum range value selection for the dynamic range. One of the advantages of this proposed method is that the use of histogram equalisation reduces sensitivity to initialisation which most thresholding algorithms suffer from. A second advantage is that it eliminates most of the background making it easy to use a simple method for candidate validation such as that in [61].

The Background Subtraction method featured a strategy for building a reliable background image without the need to use the whole video frame. The motivation for this algorithm is that the imperfections of the infrared image, specifically its low resolution and noisy nature, can become advantages. Also, as the appearance of objects in TIR footages does not change rapidly, it is enough to distort the pixels in areas where motion is detected is to obtain a background image able to extract pedestrians from the rest of the video frames.

The Candidate validation method is a semi-supervised single model for pedestrian detection. These two attributes are significant: semi-supervised and single model. First, methods in pedestrian detection are either supervised or unsupervised. Semi-supervised methods have not found much application in the thermal domain for pedestrian detection. Single model means that it combines the task of several modules into one for better results. Graph Cut [24] incorporates region and boundary into an energy function for labelling problems and is a semi-supervised method, therefore, it is adopted in this study. Also, motivated by the supervised framework in [23], the semi-supervised single-model method seeks to eliminate the need for separate modules for candidate generation and validation. Therefore, in addition to the region and boundary property, a motion constraint is incorporated into the Graph Cut energy function.

The results were presented and showed that the objectives of this research were achieved. The performance was tested on four publicly available databases. OSU thermal pedestrian database is the oldest, most comprehensive in ground truth and most extensively used, thus it allowed comparisons with other methods in the literature. LTIR is the most recent of the four databases and provided the opportunity to test the performance of the methods on modern cameras and across various scenes. It is important to note that the LTIR is not strictly for pedestrians nor pedestrian detections, therefore, selections had to be made from the database as necessary. TMIR was a straightforward database with few changes but was essential in testing the motion aspect of the proposed algorithms. LITIV images appear simple, but they are examples of long-tailed histograms which have most of their information compressed in a small area of the histogram (see Fig. 3.8).

5.2 Recommendations for Future Work

The work done in this research is part of efforts towards achieving persistent monitoring for increased security by augmenting existing video surveillance systems with thermal information. Although the differences in characteristics between visible light and thermal infrared images present different challenges for pedestrian detection, future work will involve exploring the fusion of information from both modalities for improved detections and, ultimately, a more efficient video surveillance system.

Bibliography

[1] Amanda Berg, Jörgen Ahlberg, and Michael Felsberg. A thermal Object Tracking benchmark. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2015.

[2] Mordor Intelligence LLP. Video Surveillance System Market - Growth, Trends, and Forecast (2019 - 2024). https://www.mordorintelligence.com/industry-reports/global-video-surveillance-market-industry, 08 2019.

[3] https://www.istockphoto.com/search/2/image?phrase=security+camera+pole. Accessed: 2021-09-30.

[4] C. William R. Webster. CCTV Policy in the UK: Reconsidering the Evidence Base. *Surveillance & Society*, 6(1):10–22, 2009.

[5] *Video Surveillance Market by System (Analog & IP), Offering (Hardware, Software & Service), Vertical (Commercial, Infrastructure, Military & Defense, Residential, Public Facility & Industrial), and Geography - Global Forecast to 2023.* https://www.marketsandmarkets.com/Market-Reports/video-surveillance-market-645.html, 2019. SE 2873.

[6] Mary W. Green. The Appropriate And Effective Use of Security Technologies In U.S. Schools: A Guide for Schools And Law Enforcement Agencies. *Washington, DC: U.S. Dept. of Justice, Office of Justice Programs, National Institute of Justice*, 10 1998.

[7] James Miller, Matthew Smith, and Michael McCauley. Crew Fatigue and Performance on US Coast Guard Cutters. *US Dept of Transportation*, 1999.

[8] Athanasios Tsitsoulis and Nikolaos Bourbakis. A First Stage Comparative Survey on Human Activity Recognition Methodologies. *International Journal on Artificial Intelligence Tools*, 22(06), 12 2013.

[9] Pedestrian and bicyclist detection with thermal imaging cameras. https://www.flir.eu/discover/traffic/urban/

pedestrian-and-bicyclist-detection-with-thermal-imaging-cameras/, last accessed on 30/03/2022.

[10] Nermin K. Negied, Elsayed E. Hemayed, and Magda B. Fayek. Pedestrians' detection in thermal bands – critical survey. *Journal of Electrical Systems and Information Technology*, 2(2):141 – 148, 2015.

[11] Isaac Maw. The Value of Thermal Vision in ADAS. https://www.engineering.com/AdvancedManufacturing/ArticleID/17150/The-Value-of-Thermal-Vision-in-ADAS.aspx, last accessed on 30/03/2022.

[12] Thermal Imaging vs. Night Vision Devices. https://www.opticsplanet.com/howto/how-to-thermal-imaging-vs-night-vision-devices.html, last accessed on 30/03/2022.

[13] Hot spot detection—what to look for. https://www.fluke.com/en/learn/blog/thermal-imaging/hot-spot-detection, last accessed on 30/03/2022.

[14] Michael Stuart. A Practical Guide to Emissivity in Infrared Inspections. *Uptime*, pages 43 – 46, 02 2016. Available online https://reliabilityweb.com/articles/entry/a-practical-guide-to-emissivity-in-infrared-inspections, last accessed on 30/03/2022.

[15] Amanda Berg. *Detection and Tracking in Thermal Infrared Imagery*. PhD thesis, Linkoping University, 2016.

[16] James W. Davis and Mark A. Keck. A Two-Stage Template Approach to Person Detection in Thermal Imagery. In *Proceedings of the Seventh IEEE Workshop on Applications of Computer Science*, WACV/MOTION '05, 2005.

[17] Atousa Torabi, Guillaume Masse, and Guillaume-Alexandre Bilodeau. An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications. *Computer Vision and Image Understanding*, 116(2):210–221, 2012.

[18] Emmanuel Goubet, Joseph Katz, and Fatih Porikli. Pedestrian tracking using thermal infrared imaging. In *Infrared Technology and Applications XXXII*, volume 6206, pages 797 – 808, 2006.

[19] James W. Davis and Vinay Sharma. Robust Background-Subtraction for Person Detection in Thermal Imagery. In *IEEE Workshop on Object Tracking and Classification Beyond the Visible Spectrum*, 07 2004.

[20] Mate Krišto, Marina Ivasic-Kos, and Miran Pobar. Thermal Object Detection in Difficult Weather Conditions Using YOLO. *IEEE Access*, 8:125459–125476, 2020.

[21] P. Tumas, A. Nowosielski, and A. Serackis. Pedestrian Detection in Severe Weather Conditions. *IEEE Access*, 8:62775–62784, 2020.

[22] Zongjiang Gao, Yingjun Zhang, and Yuankui Li. Extracting features from infrared images using convolutional neural networks and transfer learning. *Infrared Physics and Technology*, 105:103237, 2020.

[23] Paul Viola, Michael Jones, and Daniel Snow. Detecting pedestrians using patterns of motion and appearance. In *Proceedings Ninth IEEE International Conference on Computer Vision*, volume 2, pages 734–741, 2003.

[24] Y.Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *Proceedings of the Eighth IEEE International Conference on Computer Vision, 2001*, volume 1 of *ICCV 2001*, pages 105–112, 2001.

[25] Glen Ahearn. Cameras that See Beyond Visible Light: Inspecting the Seen and Unseen. https://www.qualitymag.com/articles/96211-cameras-that-see-beyond-visible-light-inspecting-the-seen-and-unseen, last accessed on 30/03/2022.

[26] Thermal targets. https://thermbright.com/infantry-thermal-target-training-infantry-thermal-target-training/safety-and-id-markers/, last accessed on 30/03/2022.

[27] Soundrapandiyan Rajkumar and Chandra Mouli. Adaptive Pedestrian Detection in Infrared Images using Background Subtraction and Local Thresholding. *Procedia Computer Science*, 58:706–713, 2015.

[28] Di Wu, Jihong Wang, Wei Liu, Jianzhong Cao, and Zuofeng Zhou. An effective method for human detection using far-infrared images. In *2017 First International Conference on Electronics Instrumentation Information Systems (EIIS)*, pages 1–4, 2017.

[29] Manikanta Prahlad Manda and Hi-Seok Kim. [a fast image thresholding algorithm for infrared images based on histogram approximation and circuit theory]. *Algorithms*, 13:207, 2020.

[30] Zuoyong Li, Chuancai Liu, Guanghai Liu, Xibei Yang, and Yong Cheng. Statistical thresholding method for infrared images. *Pattern Analysis and Applications*, 14:109–126, 2011.

[31] Nobuyuki Otsu. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[32] Thierry Pun. A new method for grey-level picture thresholding using the entropy of the histogram. *Signal Processing*, 2(3):223–237, 1980.

[33] Z. Hou, Q. Hu, and W. L. Nowinski. On minimum variance thresholding. *Pattern Recognition Letters*, 27(14):1732–1743, oct 2006.

[34] C.H. Li and C.K. Lee. Minimum cross entropy thresholding. *Pattern Recognition*, 26(4):617–625, 1993.

[35] Nikhil Ranjan Pal. On minimum cross-entropy thresholding. *Pattern Recognition*, 29:575–580, 1996.

[36] A.D. Brink and N.E. Pendock. Minimum cross-entropy threshold selection. *Pattern Recognition*, 29(1):179–188, 1996.

[37] Ghada Al-Osaimi and Ali El-Zaart. Minimum Cross Entropy Thresholding for SAR Images. In *2008 3rd International Conference on Information and Communication Technologies: From Theory to Applications*, pages 1–6, 2008.

[38] Tao Wu, Rui Hou, and Yixfmaiang Chen. Cloud Model-Based Method for Infrared Image Thresholding. *Mathematical Problems in Engineering*, 2016: 1–18, 2016.

[39] Jiabao Wang, Yafei Zhang, Jianjiang Lu, and Y. Li. Target Detection and Pedestrian Recognition in Infrared Images. *Journal of Computers*, 8:1050–1057, 2013.

[40] Songze Lei, Xiaoping Li, Feng Xiao, and Shifang Zhang. Pedestrian Detection Method in Infrared Images Using Maximum Entropy Threshold and Random Forest. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 10:105–114, 2017.

[41] Suman Kumar Choudhury, Pankaj Kumar Sa, Ram Prasad Padhy, Saurav Sharma, and Sambit Bakshi. Improved pedestrian detection using motion segmentation and silhouette orientation. *Multimedia Tools Application*, 77: 13075—-13114, 2018.

[42] Eun Som Jeon, Jong-Suk Choi, Ji Hoon Lee, Kwang Yong Shin, Yeong Gon Kim, Toan Thanh Le, and Kang Ryoung Park. Human Detection Based on the Generation of a Background Image by Using a Far-Infrared Light Camera. *Sensors*, pages 6763–6787, 2015.

[43] D. Jeyabharathi and Dejey. Efficient background subtraction for thermal images using reflectional symmetry pattern (RSP). *Multimedia Tools and Applications*, 77(17):22567–22586, 2018.

[44] Merzouk Younsi, Moussa Diaf, and Patrick Siarry. Automatic multiple moving humans detection and tracking in image sequences taken from a stationary thermal infrared camera. *Expert Syst. Application*, 146:113171, 2020.

[45] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[46] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-Based Visual Saliency. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, pages 545–552, 2006.

[47] Xiaodi Hou and Liqing Zhang. Saliency Detection: A Spectral Residual Approach. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[48] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597–1604, 2009.

[49] Yingfeng Cai, Ze Liu, Hai Wang, and Xiaoqiang Sun. Saliency-based pedestrian detection in far infrared images. *IEEE Access*, 5:5013–5019, 2017.

[50] [hot spot method for pedestrian detection using saliency maps, discrete chebyshev moments and support vector machine.

[51] Dahai Yu, Junwei Han, Yibo Ye, and Zhijun Fang. A novel saliency detection framework for infrared thermal images. *2014 International Conference on Orange Technologies*, pages 57–60, 2014.

[52] Rajkumar Soundrapandiyan and Chandra Mouli. Target Detection in Infrared Images Using Block-Based Approach. In *Informatics and Communication Technologies for Societal Development*, pages 9–16. Springer India, 2015.

[53] Duyoung Heo, Eun-Ju Lee, and ByoungChul Ko. Pedestrian detection at night using deep neural networks and saliency maps. *Electronic Imaging*, 2018:060403–1–060403–9, 2017.

[54] Jimei Yang and Ming-Hsuan Yang. Top-Down Visual Saliency via Joint CRF and Dictionary Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(3):576–588, 2017.

[55] Nian Liu, Junwei Han, and Ming-Hsuan Yang. PiCANet: Learning Pixel-Wise Contextual Attention for Saliency Detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3089–3098, 2018.

[56] Debasmita Ghose, Shasvat Desai, Sneha Bhattacharya, Deep Chakraborty, Madalina Fiterau, and Tauhidur Rahman. Pedestrian detection in thermal images using saliency maps. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 988–997, 2019.

[57] Yunfan Chen and Hyunchul Shin. Pedestrian Detection at Night in Infrared Images Using an Attention-Guided Encoder-Decoder Convolutional Neural Network. *Applied Sciences*, 10(3), 2020.

[58] Mai Thanh Nhat Truong and Sanghoon Kim. A study on visual saliency detection in infrared images using boolean map approach. *Journal of Information Processing Systems*, 16(5):1183–1195, 2020.

[59] A. K. M. Fahim Rahman, Mostofa Rakib Raihan, and S. M. Mohidul Islam. Pedestrian Detection in Thermal Images Using Deep Saliency Map and Instance Segmentation. *International Journal of Image, Graphics and Signal Processing*, 13:40–49, 2021.

[60] Jianming Zhang and Stan Sclaroff. Exploiting surroundedness for saliency detection: A boolean map approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5):889–902, 2016.

[61] Antonio Fernández-Caballero, María T. López, and Juan Serrano-Cuerda. Thermal-Infrared Pedestrian ROI Extraction through Thermal and Motion Information Fusion. *Sensors*, 14(4):6666–6676, 2014.

[62] Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. In *2005 IEEE Conference Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[63] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144—-152, 1992.

[64] Wei Li, Dequan Zheng, Tiejun Zhao, and Mengda Yang. An effective approach to pedestrian detection in thermal imagery. In *2012 8th International Conference on Natural Computation*, pages 325–329, 05 2012.

[65] Soundrapandiyan Rajkumar and Chandra P. Mouli. An Approach to Adaptive Pedestrian Detection and Classification in Infrared Images Based on Human Visual Mechanism and Support Vector Machine. *Arabian Journal for Science and Engineering*, 43:3951–3963, 2018.

[66] Jeonghyun Baek, Sungjun Hong, Jisu Kim, and Euntai Kim. Efficient Pedestrian Detection at Nighttime Using a Thermal Camera. *Sensors (Basel, Switzerland)*, 17(8), 2017.

[67] Zelin Li, Wu Qiang, Jian Zhang, and Glenn Geers. SKRWM based descriptor for pedestrian detection in thermal images. In *2011 IEEE 13th International Workshop on Multimedia Signal Processing*, pages 1–6, 2011.

[68] Jisoo Park, Jingdao Chen, Yong K. Cho, Dae Y. Kang, and Byung J. Son. CNN-Based Person Detection Using Infrared Images for Night-Time Intrusion Warning Systems. *Sensors*, 20(34), 2019.

[69] Noor Ul Huda, Bolette D. Hansen, Rikke Gade, and Thomas B. Moeslund. The Effect of a Diverse Dataset for Transfer Learning in Thermal Person Detection. *Sensors*, 20(7), 2020.

[70] Kieu My, Lorenzo Berlincioni, Leonardo Galteri, Marco Bertini, Andrew Bagdanov, and Alberto Bimbo. Robust pedestrian detection in thermal imagery using synthesized images. In *25th International Conference on Pattern Recognition, ICPR 2020*, pages 8804–8811, 2020.

[71] Shasha LI, Yongjun Li, Yao Li, Mengjun Li, and Xiaorong Xu. YOLO-FIRI: Improved YOLOv5 for Infrared Image Object Detection. *IEEE Access*, 9: 141861–141875, 2021.

[72] Roland Miezianko. IEEE OTCBVS WS Series Bench; Terravic Research Infrared Database. http://vcipl-okstate.org/pbvs/bench/Data/05/download.html, last accessed on 31/03/2022.

[73] Bo Lei and Jiulun Fan. Multilevel minimum cross entropy thresholding: A comparative study. *Applied Soft Computing*, 96:106588, 2020.

[74] D. Coltuc, P. Bolon, and J.-M. Chassery. Exact histogram specification. *IEEE Transactions on Image Processing*, 15(5):1143–1152, 2006.

[75] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[76] Claude E. Shannon and Warren Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1949.

[77] J. Shore and R. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, 26(1):26–37, 1980.

[78] Stuart Geman and Donald Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[79] Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society Series B*, 48(3):259–302, 1986.

[80] D. M. Greig, B. T. Porteous, and A. H. Seheult. Exact Maximum A Posteriori Estimation for Binary Images. *Journal of the Royal Statistical Society. Series B (Methodological)*, 51(2):271–279, 1989.

[81] L. R. Ford and D. R. Fulkerson. Maximal flow through a network. *Canadian Journal of Mathematics*, 8:399–404, 1956.

[82] Kamal Hajari Ujwalla Gawande and Yogesh Golhar. Pedestrian detection and tracking in video surveillance system: Issues, comprehensive review, and challenges. In *Recent Trends in Computational Intelligence*. IntechOpen, 2020.

[83] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004.

[84] Oluwakorede M. Oluyide, Jules-Raymond Tapamo, and Tom Walingo. [fast background subtraction and graph cut for thermal pedestrian detection.

[85] Yifan Zhao, Jingchun Cheng, W. Zhou, C. Zhang, and Xiong Pan. Infrared Pedestrian Detection with Converted Temperature Map. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 2025–2031, 2019.

[86] Manikanta Prahlad Manda, ChanSu Park, ByeongCheol Oh, Daijoon Hyun, and Hi Seok Kim. Pedestrian Detection in Infrared Thermal Images Based on Raised Cosine Distribution. In *2020 International SoC Design Conference (ISOCC)*, pages 278–279, 2020.

[87] Ali Haider, Furqan Shaukat, and Junaid Mir. Human detection in aerial thermal imaging using a fully convolutional regression network. *Infrared Physics & Technology*, 116:103796, 2021.