Methods for multi-spectral image fusion

Identifying stable and repeatable information across the visible and infrared spectra

Presented by:
Francois Jacques Retief

Prepared for:
Dr. F. Nicolls
Dept. of Electrical Engineering
University of Cape Town

Submitted to the Department of Electrical Engineering at the University of Cape Town in fulfilment of the academic requirements for a Master of Science in Engineering

April 20, 2016

Declaration

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.
2. I have used the IEEE convention for citation and referencing. Each contribution to, and quotation in, this report from the work(s) of other people has been attributed, and has been cited and referenced.
3. This report is my own work.
4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as their own work or part thereof.

Signature: 

F. J. Retief

Date:30 April 2016

Acknowledgments

I would like to thank my supervisors, Dr. F. Nicolls and Dr. G. de Jager, for their valuable guidance and patience over the course of this dissertation. Further thanks go to the academic and administrative staff of the University of Cape Town for the support provided to me while on (and off) campus. Finally, I would like to express my gratitude for the financial support provided by the National Research Foundation for this project.

Abstract

Fusion of images captured from different viewpoints is a well-known challenge in computer vision with many established approaches and applications; however, if the observations are captured by sensors also separated by wavelength, this challenge is compounded significantly. This dissertation presents an investigation into the fusion of visible and thermal image information from two front-facing sensors mounted side-by-side. The primary focus of this work is the development of methods that enable us to map and overlay multi-spectral information; the goal is to establish a combined image in which each pixel contains both colour and thermal information.

Pixel-level fusion of these distinct modalities is approached using computational stereo methods; the focus is on the viewpoint alignment and correspondence search/matching stages of processing. Frequency domain analysis is performed using a method called phase congruency. An extensive investigation of this method is carried out with two major objectives: to identify predictable relationships between the elements extracted from each modality, and to establish a stable representation of the common information captured by both sensors. Phase congruency is shown to be a stable edge detector and repeatable spatial similarity measure for multi-spectral information; this result forms the basis for the methods developed in the subsequent chapters of this work.

The feasibility of automatic alignment with sparse feature-correspondence methods is investigated. It is found that conventional methods fail to match inter-spectrum correspondences, motivating the development of an edge orientation histogram (EOH) descriptor which incorporates elements of the phase congruency process.

A cost function, which incorporates the outputs of the phase congruency process and the mutual information similarity measure, is developed for computational stereo correspondence matching. An evaluation of the proposed cost function shows it to be an effective similarity measure for multi-spectral information.

Contents

Glossary

ADAS Advanced Driver Assistance Systems. 12

AGAST Adaptive and Generic Accelerated Segment Test. 61

AGC Automatic gain control. 18

AOI Automated visual inspection. 8

AUC Area under curve. 46, 92

BFROST Binary Features from Robust Orientation Segment Tests. 61

BRIEF Binary Robust Independent Elementary Feature. 63

BRISK Binary Robust Invariant Scalable Keypoint. 61, 63, 68, 69

CCD Charge-coupled device. 13

CMOS Complementary metal-oxide semiconductor. 12

CPU Central processing unit. 93

CT Computed tomography. 8, 9

CVC Computer Vision Center. 12, 13, 70

DoG Difference of Gaussian. 61

EOH Edge Orientation Histogram. 64–72, 92, 93

FAST Features from Accelerated Segment Test. 61, 67, 68, 70

FFC Flat-field correction. 16

FLIR Forward-looking infrared. 3

FN False negative. 46

FP False positive. 46

FPA Focal plane array. 16

FPR False positive rate. 46

FREAK Fast Retina Keypoint. 63

GDI-SIFT Gradient Direction Invariant SIFT. 63

GF Good Features. 61, 68–71, 92, 93

GI Gradient information. 82, 83, 87, 92

GPU Graphics processing unit. 93

HOG Histogram of Oriented Gradients. 64

I Intensity information. 82, 83, 87

LWIR Longwave infrared. 11

MGI The product of mutual information (MI) and gradient information (GI) measures.. 82–85, 87, 93

MI Mutual information. 82, 83, 85, 87, 92, 93

MRI Magnetic resonance imaging. 8, 9

NIR Near infrared. 11, 12

NMS Non-maximal suppression. 80

OCTBVS Object Tracking and Classification in and Beyond the Visible Spectrum. 13, 14

ORB Oriented Fast and Rotated BRIEF. 61, 63

OR-SIFT Oriented SIFT. 63

OSU Oklahoma State University. 13, 14, 48, 51, 69, 83, 84, 86, 88, 89

PC Phase Congruency edge information. 82, 83, 87, 88, 93

PCB Printed circuit board. 8

PET Positron emission tomography. 8

PSU Power supply unit. 8, 17

RANSAC Random Sample And Consensus. 68, 70

RBF Radial Basis Function. 44, 45, 54

ROC Receiver Operating Characteristic. 45, 46, 52, 54, 92

SIFT Scale invariant feature transform. 61–63, 65, 67–70

SMD Surface-mounted device. 8

SURF Speeded Up Robust Features. 62, 63, 67–70

SVM Support Vector Machine. 44–47, 91, 92

SWIR Shortwave infrared. 11

TN True negative. 45

TP True positive. 45

TPR True positive rate. 46, 82–85, 87

UR-SIFT Uniform Robust SIFT. 63

USB Universal serial bus. 3, 17

VGA Video graphics adapter. 3

VS Visible spectrum. 11

Y Normalised mutual information. 82, 85

YGI The product of normalised mutual information (Y) and gradient information (GI). 82, 84, 87

Chapter 1

Introduction

1.1 Subject and aims

Within the past decade, imaging technology capable of capturing information beyond the visible spectrum has become widely available. These devices operate at different spectral ranges to capture distinct information about a scene. The strengths of each spectral modality can be simultaneously utilised through a process called information fusion. The primary objective of this dissertation is to establish a set of methods that facilitate multi-spectral image fusion.

Approaches to this task are numerous and varied, but largely operate on the same assumptions about the characteristic behaviour of illumination in the scene; however, if the observations occur in different spectral ranges, these assumptions quickly fall apart. This dissertation presents a series of investigations to identify, adapt, develop and evaluate the methods required to fuse the disparate information captured by thermal and visible spectrum sensors.

1.2 Motivation for study

Visible spectrum cameras have become embedded in our everyday life and provide high quality images at a low price. Broad application domains such as surveillance,

process control and automated visual inspection rely on the information provided by imaging systems to perform complex tasks. However, deployment of these systems is limited by their reliance on scene illumination or the presence of occluding atmospheric conditions (e.g. fog or smoke).

Thermal scenes are independent of scene illumination and capture a very different representation of the world from the one we are familiar with. Infrared radiation is emitted by any object at a temperature above absolute zero; the amount of radiation emitted is based on the material and temperature. Thermal imaging is still young as a consumer product, and prices remain very high and scale disproportionately with sensor resolution. This high cost for spatial resolution is a major limiting factor in the adoption of thermal imaging.

The distinct spectral information captured by thermal and visible sensors lends each to excel at different applications. What is interesting is the degree to which the strengths and functionality provided by these two modalities complement each other.

Thermal images appear smooth and textureless due to a lack of colour information and shadowing — two important visual cues we rely on to perceive texture. Consumer-grade thermal cameras typically have a very low spatial resolution (e.g. 320×240 pixels is common) which further reduces the clarity of contours and edges. Visible spectrum images provide increased edge fidelity at a higher resolution, and are therefore a cost effective way of enhancing the performance of thermal imaging devices through fusion. Furthermore, the perception of colour and texture provided by the visible spectrum is essential in correspondence and recognition applications.

Thermal imaging is particularly useful in applications which require stable 24-hour operation in highly variable illumination (e.g. day/night cycles, moving shadows) or occluding atmospheric (e.g. fog or smoke) conditions. Additionally, the use of thermal information lends valuable functionality and robustness to detecting, isolating and tracking objects against cluttered backgrounds. The relationship between material emissivity and observed thermal intensity means that image segmentation can be aided by the inclusion of thermal information in providing material boundaries.

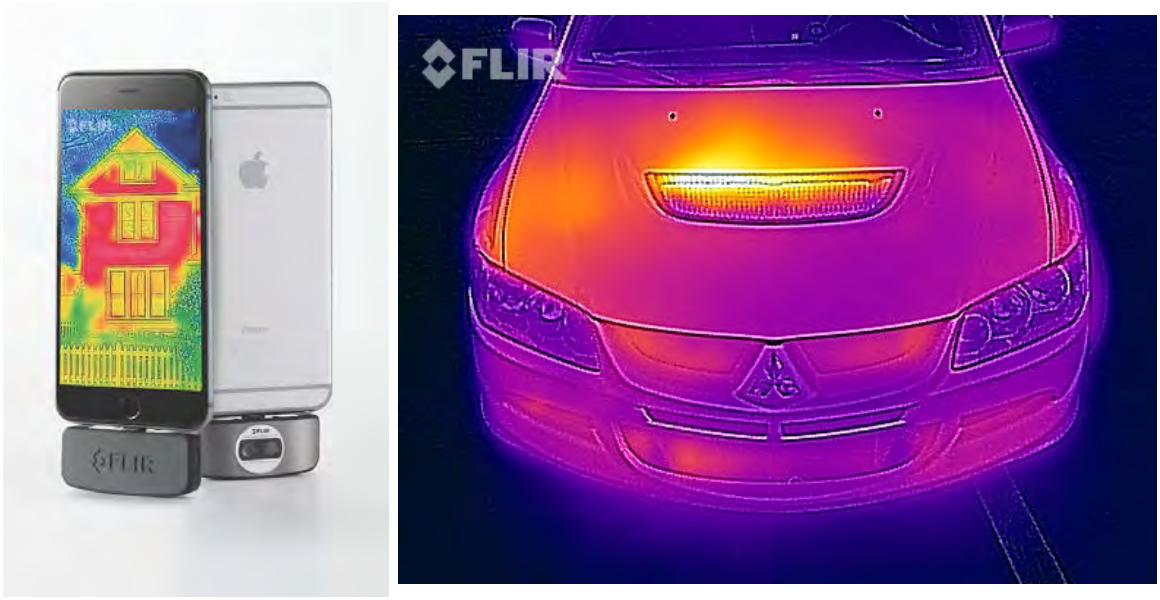
Fusion of thermal and visible information provides complementary functionality which is often required in many generic vision-based systems. For example, military and surveillance systems require target detection (thermal) followed by identification (visible); additionally, autonomous machines can benefit from this enhanced scene

understanding for more complex or robust decision making. Inexpensive thermal cameras do exist, but have a very low spatial resolution (e.g. 80×60 pixels); however, a higher spatial resolution for the thermal camera can be emulated through fusion with a regular camera. The FLIR ONE, presented in the next section, achieves this to provide functional thermal imaging in the form of a small, inexpensive mobile phone add-on.

FLIR ONE: A case study

The FLIR ONE is a $249 micro-Universal serial bus (USB) add-on for Android and iOS smartphones. The attachment is an example of multi-spectral fusion to utilise the resolution and clarity of a visible spectrum camera to enable the practical use of an inexpensive thermal camera with a very low resolution.

Shown in Fig. 1.1[1], the micro-USB add-on, which weighs only 110g, houses a Video graphics adapter (VGA) resolution visible spectrum camera and a FLIR Lepton thermal micro-core separated by 20mm. The 80×60 pixel thermal image produced by



(a) FLIR ONE iPhone add-on. (b) The low resolution thermal information is interpolated onto an edge map.

Figure 1.1: The FLIR ONE device with an accompanying example image.

the uncooled Forward-looking infrared (FLIR) core is interpolated onto an "embossed"

[1]Images from product website at `http://www.flir.com/flirone`, 23 August 2015.

640×480 pixel edge map generated by the visible spectrum camera. While the camera configuration and function provided by the device are similar to the goals of this work, the investigations covered here are aimed at the fusion of more advanced thermal cameras. The incorporation of high resolution cameras and larger lenses significantly increases the complexity of the system as the distance between the apertures is no longer negligible.

1.3 Objectives and context

The aim of this dissertation is to develop methods that enable us to fuse information captured by imaging sensors operating in distinct spectral bands. Although information fusion is not a specific objective of this work, the methods that are developed form the basis for fusion algorithms. The goal is to create a single enriched image in which each pixel contains both colour and thermal information by mapping the spectral information of each point from one viewpoint to the other.

There are two tasks which are focus of this work: the alignment of the two distinct viewpoints, and the matching of corresponding regions observed by each spectral sensor. These tasks allow us to map and overlay the spectral information captured by each camera; however, both of these tasks require a means of comparing multi-spectral information to gauge similarity between regions. Central to this work is the investigation and development of a method for extracting the shared inter-spectrum information that can be used to compare regions of multi-spectral images. The extracted information must be common across the infrared and visible spectra and also remain stable under the variations present within each modality.

The first task is the alignment of the distinct viewpoints captured by the thermal and visible sensors. Alignment is a frequent task in stereo-vision systems with many established approaches; therefore, the performance of traditional methods for automatic alignment using feature-based methods is investigated. The aims of the investigation are to determine the suitability of existing methods, and to evaluate adaptations for multi-spectral feature point matching.

The second task is to develop an approach to mapping and overlaying corresponding regions of multi-spectral information. For this objective, it will be assumed that the images have been correctly aligned to make the problem more tractable. The

requirements of the method are dependent on the challenges associated with thermal information and the camera configuration, and the implementation will incorporate the multi-spectral similarity measure developed in this work.

1.4 Outline

Chapters 2 and 3 provide a background to multi-spectral imaging and clarify the objectives and approach to fusion. Subsequent chapters are structured as a series of investigations addressing specific development and analysis tasks laid out in these preliminary chapters.

Chapter 2 presents a brief literature review to establish the context of this work in both multi-spectral imaging and information fusion research. As the majority of research on visible and infrared imaging is focused on specific applications, relevant research is grouped and presented according to the approach to fusion. A small set of publications which particularly influenced the approach to development in later chapters is discussed, and the datasets available and in use at the time of writing are provided.

Chapter 3 reviews aspects of thermal imaging and provides clarity on the goals and approach taken in this work towards fusion. The camera configuration is specified; the dual-camera set-up and its relation to the scene dictate the steps that need to be taken to fuse the distinct viewpoints. First, calibration of multi-spectral vision systems and the requirements for alignment are discussed. Computational stereo is then introduced as the means of overlaying (or mapping) the information from the two viewpoints.

Chapters 4, 5 and 6 present three structured investigations. The theory and approach taken to accomplishing each objective is detailed in the relevant chapter, and the results obtained in each investigation are used to guide development in subsequent chapters. The approach to fusion is divided into two steps: alignment and correspondence mapping, which are addressed in Chapters 5 and 6 respectively.

Chapter 4 introduces a frequency-domain analysis tool called phase congruency, which is the proposed method for identifying and exposing the stable features common to both visible and thermal images. There are two goals to this chapter: to expose repeatable structural features (i.e. edges and corners), and to identify predictable

components extracted by the phase congruency process.

Repeatable structural features provide an invariant representation of the scene. This representation transforms the multi-spectral images to appear visually similar by removing the varying elements (e.g. brightness or contrast). The second goal is aimed at identifying a predictable relationship between points observed from the two spectral modalities, which would provide a valuable means of comparing and matching corresponding points; identifying and evaluating such features is the focus of the analysis carried out in Chapter 4.

Chapter 5 presents an investigation into methods for the automatic alignment of multi-spectral images. Conventional feature-based methods detect and match significant points in each image to bring them into alignment. An adapted method, which incorporates the invariant representation provided by phase congruency, is proposed and benchmarked against traditional methods. The aim of this chapter is to determine the feasibility of automatic alignment of multi-spectral images using current (and adapted) feature-based methods.

Chapter 6 presents the development of a cost function for computational stereo to measure the similarity of regions in the search for corresponding points in multi-spectral image pairs. Once again, the repeatable features extracted with the phase congruency process are integral to the similarity measure and approach to computational stereo.

Chapter 7 provides an overview on the findings of this project and proposes extensions for future work.

Chapter 2

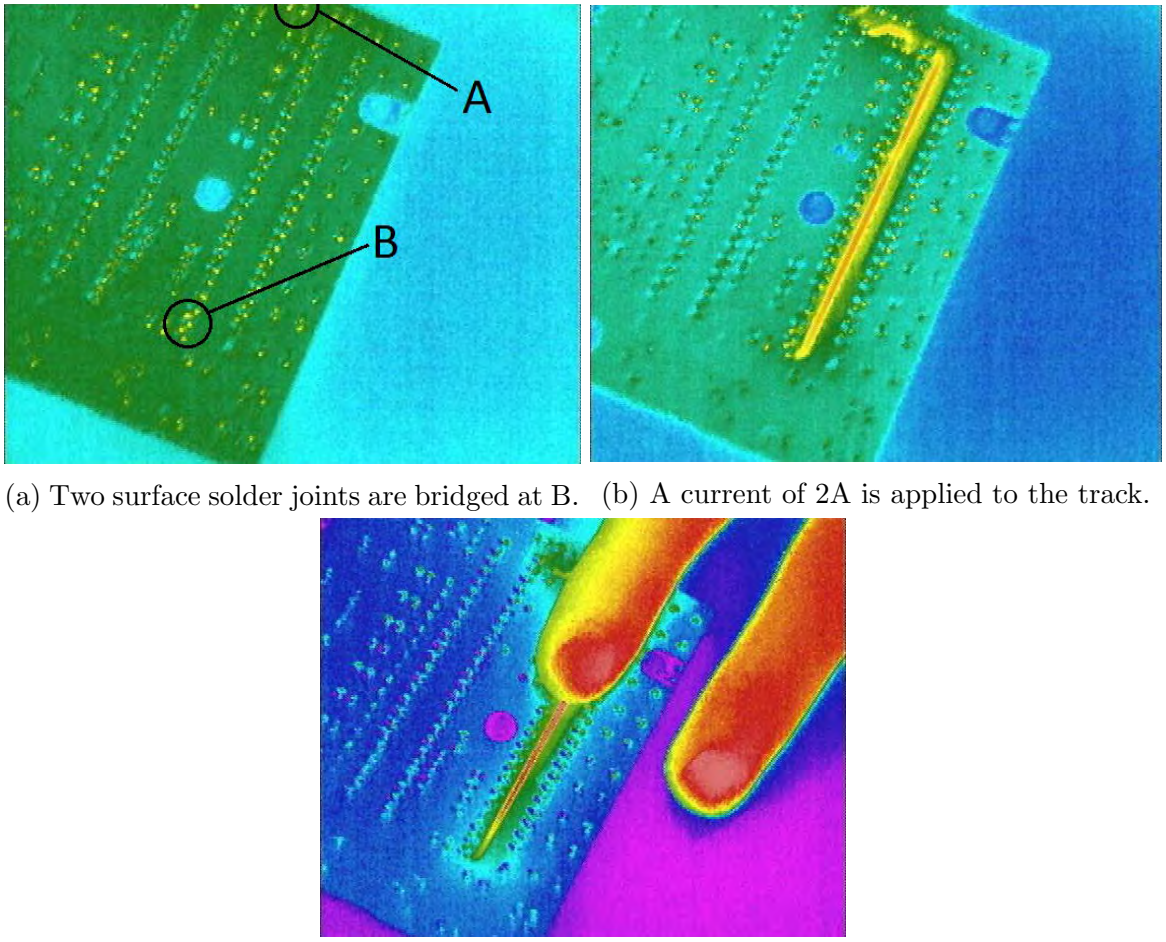
Literature Review

This chapter provides a background to multi-spectral imaging and fusion systems. A broad overview of multi-spectral image systems is presented in Section 2.1, followed by an introduction to the classes of image fusion in Section 2.2, to establish the terminology used to describe the undertaken approach to fusion in the context of literature. Section 2.3 introduces infrared radiation and, using the established classification of fusion systems, provides a brief taxonomy of visible-infrared fusion systems. Finally, contemporary research that closely influenced the content of this work is presented, with examples of the available visible-infrared datasets and where they are used.

2.1 Imaging beyond the visible spectrum

Multi-sensor image fusion is a broad field of study that dates back to the 1950s [1, 2]. Initial deployment of spectral imaging systems began in the 1960s on satellite and airborne hyper-spectral platforms capable of capturing hundreds of spectral bands. This high spectral resolution was of particular benefit in applications including the detailed analysis of crop conditions, mineral exploration and for monitoring the effects of urban development [3]. On the other hand, *multi*-spectral imaging typically involves the simultaneous capture of 2 to 6 spectral bands; the focus is placed on smaller systems in real-time applications and non-contact analysis [2].

Multi-spectral imaging is often integrated into manufacturing and quality control processes. Printed circuit board (PCB) quality control, termed Automated visual inspection (AOI) systems, incorporate x-ray, ultrasonic as well as thermal imaging [4]. AOI systems are often able to pick up common defects in PCBs that are missed by contact testing rigs. For example, x-ray AOI is able to detect defects in multi-layered PCBs such as hairline cracks, Surface-mounted device (SMD) misalignment, track bridging or incorrectly soldered components [4, 5]. Thermal imaging, although harder to automate, is able to detect hot-spots due to short-circuits and faulty or stressed components. An example of a short-circuited track on a PCB is shown in Fig. 2.1.



(a) Two surface solder joints are bridged at B. (b) A current of 2A is applied to the track.

(c) The tracks remain cool to the touch.

Figure 2.1: A short circuit (bridge) is created at point B. Two parallel copper tracks that run from A to B are visible under thermal imaging as $2A$ is applied using a current-controlled power supply unit (PSU). The heat emitted by the tracks is due to the resistance of the narrow copper tracks.

Medical imaging technologies such as x-ray, Computed tomography (CT), Magnetic resonance imaging (MRI) and Positron emission tomography (PET) scans are useful non-invasive diagnostic tools that have greatly benefited from combined use through

multi-spectral fusion [6]. Different structures in the human body are optimally captured in different spectral ranges — prompting the combined use of this specialised spectral information to form a diagnosis and plan treatment. An example of this is found in radiotherapy where MRI information is used to outline tumours and a CT scan is used to provide the tissue information required to calculate the dose [7]. The fusion of these modalities has become well researched over the past decade; Pluim et al. provide an extensive taxonomy on the field [8].

2.2 Information fusion

The aim of information fusion is to combine multiple heterogeneous sources of complementary information into a more useful representation; however, what is considered useful information is determined by requirements of the specific application [1, 9]. For example, if the representation is intended for human operators, the information presented must be relevant to the task at hand and minimise ambiguity. Alternatively, in the context of automated systems, information fusion is used to enhance scene understanding or increase reliability in feature extraction on huge datasets [1, 10–12].

Typical fusion systems are categorised into four classes; this categorisation is based on the amount of processing that has been performed on the input before it is fused [2]. The four classes of fusion are:

- *Observation level*: The raw input of sensors measuring the same quantity is fused through a simple operation, based on a known relationship between sensors [10] (e.g. multiple microphones measuring acoustic information [13]). Fusion of the redundant information captured by this type of system produces a signal of the same form with increased fidelity, reduced uncertainty and robustness against sensor failure [14].

- *Pixel level*: The amount of information at each pixel is increased by the fusion of multiple images. Typically, the cooperative use of multiple viewpoints (in the same spectral range) is used to recover depth information through the known relationship between the two apertures. The fusion of multiple observations of a scene has also been shown to improve the performance of traditional image processing tasks like segmentation and feature extraction [10, 14].

- *Feature level*: Features containing salient information are extracted from the sensor data. This level of fusion is geared towards large-scale analysis: feature vectors are combined or concatenated into distinctive vectors so that machine learning or pattern recognition methods can be more effective [13].

- *Decision level*: This high-level fusion step is done after processing the sensor information [1]. The aim is to enhance the understanding of objects in the scene [10]. Inferences made about objects and object behaviour in a setting (e.g. pedestrians, obstacles) can be significantly improved with the added information. Visual representation can also be more nuanced, based on rule-based semantics to aid interpretation of a scene.

Application-driven literature will often accomplish feature and decision-level fusion based on prior knowledge of scene elements (e.g. the detection and tracking of pedestrians) or specific objectives (e.g. spectral feature analysis on globally aligned aerial photography). This work is focused on pixel-level fusion of thermal and visible spectral information. As this class of fusion occurs at a very low level, just after the images have been captured, the solution does not rely on specific assumptions about the scene; therefore, the methods developed in this work can be applied to a wide array of applications.

2.3 Applications of infrared radiation

This section focuses on infrared radiation and its applications in fusion. A brief background to the electromagnetic spectrum is provided in Section 2.3.1, followed by a taxonomy of infrared-visible fusion systems grouped into observation/pixel-level and feature/decision-level classes of fusion in Sections 2.3.2 and 2.3.3 respectively.

2.3.1 Infrared radiation

In 1800, Frederick William Hershel (1738–1822), known for his numerous astronomical discoveries and musical accomplishments, conducted experiments to measure the energy in each colour of the spectrum. Hershel used a glass prism to separate the wavelengths of light onto a series of thermometers placed at each spectral band. He

noticed a significant amount of heat detected beyond the red spectral band which behaved like visible light — he had discovered infrared radiation. [15, 16].

The electromagnetic spectrum is a classification of radiation based on wavelength. The familiar visual spectrum lies from $0.39\mu m$ to $0.77\mu m$; this range contains the light visible to the human eye, from violet at 0.390-0.453μm to red at 0.622-0.770μm [17]. There are four classes of infrared imaging technology which operate in the $0.75\mu m$ to $15\mu m$ range [18]. The respective wavelengths of the Visible spectrum (VS), Near infrared (NIR), Shortwave infrared (SWIR), Midwave infrared (SWIR) and Longwave infrared (LWIR) or thermal bands are shown in Fig. 2.2.

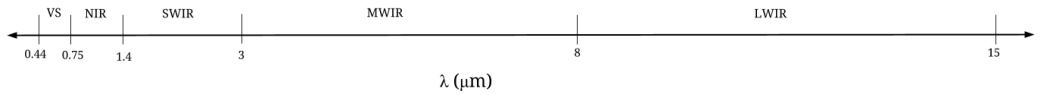


Figure 2.2: This segment of the electromagnetic spectrum shows the wavelengths of the visible and infrared spectral bands.

Sensors operating in each infrared spectral range provide different information and have been used in a wealth of applications. However, the distinct information of each spectral range and application-specific assumptions of each implementation makes it difficult to consolidate the methods into a generalised approach to working with infrared images. To address this, a brief overview of relevant research into visible-infrared fusion is presented in the next two sections; there are also helpful taxonomies and overviews available [1, 10, 13].

2.3.2 Observation/pixel-level fusion

Near infrared and short wave infrared are used in image dehazing. Haze is the loss of contrast and detail due to atmospheric conditions such as fog or pollution, and is particularly prevalent in aerial or landscape photography [19].

Near infrared is of particular interest due to the possibility of single-sensor integration. Silicon-based photosites used in consumer digital cameras are sensitive to wavelengths from 400 to 1100nm; however, a hot mirror filter is used in colour photography to block the near infrared wavelengths beyond 700nm [19, 20]. Proposed by Fattal et al. is the integration of visible and NIR receptors on a single sensor plane to allow for automatic dehazing of images taken with existing sensor technology [21]. The degree

of scattering caused by mist or haze is estimated using the NIR sensor and is used to sharpen and increase the colour fidelity of the effected regions in the visible image. Modern Complementary metal-oxide semiconductor (CMOS) sensors are increasing in photosite density due to the demand for embedded cameras [22], so hybrid sensors with different types of photosites would be able to match the image quality of regular sensors.

2.3.3 Feature/decision-level fusion

Long wave infrared (thermal) images are independent of scene illumination, and are therefore the natural choice for surveillance applications where systems must contend with scene clutter, ambient illumination (i.e. time of day) and challenging atmospheric conditions [23]. Of particular interest in visible-thermal surveillance literature is pedestrian detection and tracking in natural scenes where the use of thermal information can be used to discern human silhouettes from cooler surroundings [24–26].

In isolation, thermal imaging is unable to distinguish between similar silhouettes or objects at similar temperatures as it lacks the distinctive colour information and edge fidelity of the visible spectrum [27]. Silhouettes of pedestrians are altered by clothing, particularly in the cases of dresses and burkas [24]; however, as is common practice with existing systems, it is relatively easy to train a classifier to recognise (and subsequently track) pedestrian silhouettes [28, 29]. Related to pedestrian tracking is the cooperative use of thermal and visible imaging information in facial recognition and disguise detection, which has also garnered academic interest [30–33].

2.4 Previous work

While the approaches discussed in the previous section address the fusion of visible and thermal information, the solutions are based on assumptions about scene content (e.g. pedestrians distinct from a static background). It is difficult to discern a generalised approach to the fusion from these works, although the methods developed do provide significant hints and guidance towards such a goal. A significant amount of research has been published by members of the Advanced Driver Assistance Systems (ADAS) Computer Vision Center (CVC) at the Universitat Autonoma de Barcelona in the last

five years [2, 18, 34–38] which is more in line with the general fusion goals of this work.

Chapters 4, 5 and 6 of this work are structured as separate investigations, and the relevant literature pertaining to specific aspects in the development of the theory is presented in a brief background section to each investigation. Significant attention is paid to a frequency domain analysis method called phase congruency in Chapter 4, based on the extensive work and implementation of Kovesi [39–42]. The feature descriptor, developed in Chapter 5, is based on the work of Aguilera et al. [35] and Mouats et al. [43]. The approach of Barrera [2] was particularly influential in the cost function for computational stereo developed in Chapter 6.

There is a very narrow range of visible-thermal multimodal datasets available with ground truth disparity for automated testing [2]. There are two main datasets in use (the works that use them are cited): the CVC Multimodal Stereo Dataset[1] (1) [2] and (2) [34, 35] shown in Fig. 2.3, and the Oklahoma State University (OSU) Object Tracking and Classification in and Beyond the Visible Spectrum (OCTBVS) Color-Thermal Database[2] [44] shown in Fig. 2.4.

Figure 2.3: CVC Multimodal Stereo Dataset (2). The hand-rectified images are taken from two rigidly mounted cameras. The Sony Charge-coupled device (CCD) camera and the PathFindIR FLIR camera produce 640×480 and 534×426 pixel images respectively. Once the images have been rectified they are both 506×408 [34, 35].

Note that the OSU database (Fig. 2.3) is the only dataset with ground truth disparity information that can be used for automated analysis. Images from the CVC dataset were used, but the tests could not be automated to the same degree. Attempts have been made by authors to synthetically alter established colour datasets to overcome

[1] Available at `http://www.cvc.uab.es/adas/projects/simeve/`, August 2015.
[2] Available at `http://vcipl-okstate.org/pbvs/bench/`, August 2015.

Figure 2.4: The OSU (OCTBVS) Color-Thermal Database are a set of hand-rectified 320×240 pixel images with a ground truth disparity of ± 2 pixels [44].

this limitation. For example, the extensive Middlebury stereo dataset[3], which is the standard for computational stereo methods [45], is altered by introducing non-linear intensity variations to regions in the image pairs [46]. However, methods developed in this way have been shown to not generalise to natural scenes captured by the sensors themselves [47].

In summary, this chapter aims to establish the context of this work in multi-spectral imaging and fusion literature; the overarching goal describing this work is identified as pixel-level fusion of thermal and visible spectrum images. The next chapter introduces the characteristics (and significant challenges) of the thermal spectrum, and the approach and specific objectives addressed in this work are clarified.

[3]Available at `http://vision.middlebury.edu/mview/`, August 2015.

Chapter 3

Background to vision systems

The purpose of this chapter is to identify the methods required for multi-spectral image fusion. Theory and concepts of stereo-vision systems are presented to establish the specific objectives for the investigations carried out in later chapters of this work. Section 3.1 describes thermal imaging technology and the characteristics of thermal scenes to introduce the core challenges of visible-infrared fusion. The camera configuration and approaches to computational stereo are introduced in Section 3.2, with a focus placed on calibration and alignment of multi-sensor systems and methods for region-based matching.

3.1 Imaging technology

Thermal information is very different from the familiar visible spectrum; the characteristics and challenges of thermal imaging are described in this section. These challenges form the basis for the development work carried out in Chapter 4. There are two parts to this section. As thermal imaging is still a niche topic, a brief background to thermal radiation and image capture is first provided, followed by an overview of the challenges associated with the behaviour of objects and materials in thermal scenes.

3.1.1 Thermal image formation

Digital cameras have become embedded in much of the technology we use on a daily basis; from cellphone cameras to city-wide surveillance networks, this versatile technology is commonplace and widespread. The ever increasing demand for digital cameras has led to the rapid development of the technology to meet consumer expectations of image quality and to conform to size, weight and power constraints of mobile devices [48, 49].

Thermal imaging technology is only just beginning to be introduced into mainstream markets and is yet to gain traction. The image capture process of a thermal camera is very similar to that of a common digital camera. Modern digital camera sensors contain millions of monochromatic photosites[1] which react to the intensity of the incident visible light. Colour is obtained by superimposing a colour filter of juxtaposed red, green and blue filters onto the sensor plane [49]; the colour arrangement on these band-pass filters commonly follows the Bayer Pattern [50]. Thermal lenses are composed of multiple germanium layers that act as band-pass filters and focus infrared radiation onto the IR-sensitive photosites (e.g. bolometers) of the Focal plane array (FPA) [2].

Consumer-grade thermal cameras are uncooled and of a similar size and weight to modern digital cameras; an example can be seen in Fig. 3.1, which shows the thermal camera used to demonstrate typical thermal phenomena in later sections of this work.

Thermal cameras of this class typically have a low spatial resolution which causes high-frequency information (e.g. fine-grained detail and edges) to be lost [43]. The uncooled cores are prone to high levels of noise caused by fluctuations in ambient temperature and the temperature of the device itself, which manifests as graininess in the image. The re-calibration process is done by applying periodic Flat-field correction (FFC) updates in which a shutter of uniform temperature is applied to every thermal photosite to establish a uniform thermal baseline on the FPA.

The range (distance) of a thermal camera is dependent on the focal length of the lens and the atmospheric conditions. The performance of a thermal lens is commonly specified in terms of the maximum distances of detection, recognition and identification of a human. A modern FLIR high definition camera (HDC)[2] camera is able to detect

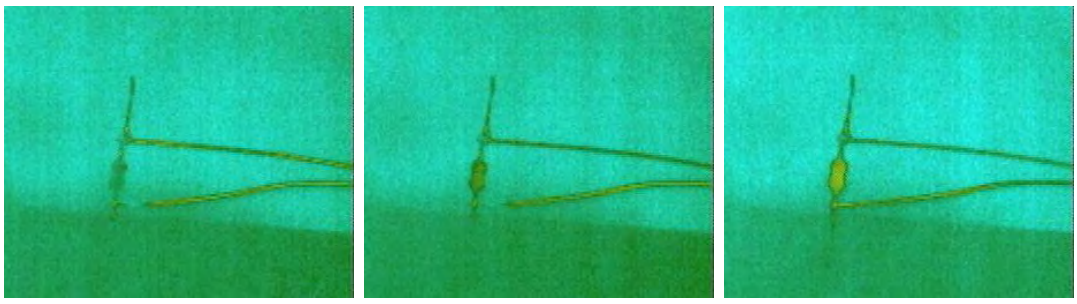
[1] A term used by Mancuso and Battiato [49] to describe the photo-sensitive diodes on the sensor.
[2] Information available at `http://www.flir.co.uk/cs/display/?id=60097`, August 2015.

(a) The Tau module. **A** shows the analogue video output micro coaxia cable; the mini-USB port powers the device.

(b) The Tau320-P is mounted on a small tripod. The 19mm lens has a detection range of 450 meters.

Figure 3.1: The Tau320-P camera provides 320×240 pixel infrared ($7.5 - 13.5$ μm) images at 9 frames per second (slowed for US export).

humans up to 18km away and vehicles up to 22km away. The thermal resolution of images is quantised based on the range of temperatures present in the scene. Figure 3.2 shows an experiment carried out with the Tau320-P camera to demonstrate the thermal sensitivity of the sensor based on the power dissipated by a 150Ω resistor.



(a) Ground truth (ambient temperature).

(b) 5mW is disappated by the resistor.

(c) 17mW is disappated by the resistor.

Figure 3.2: Thermal images of a 150Ω resistor demonstrate the thermal sensitivity of the Tau320-P by capturing the heat dissipated when a current is passed through the resistor using a current-controlled Power supply unit (PSU).

3.1.2 Thermal phenomena

The information captured by thermal sensors is distinct from the familiar visible spectrum where colour and texture is used to describe objects. The infrared intensity radiated by an object is a product of its temperature and the characteristic emissivity and reflectance of its constituent materials and surface finish [23,51]. Figure 3.3 shows a series of images of a Raspberry Pi[3] mini-computer booting up; the surface mounted components radiate and diffuse thermal radiation and illuminate the circuit board.

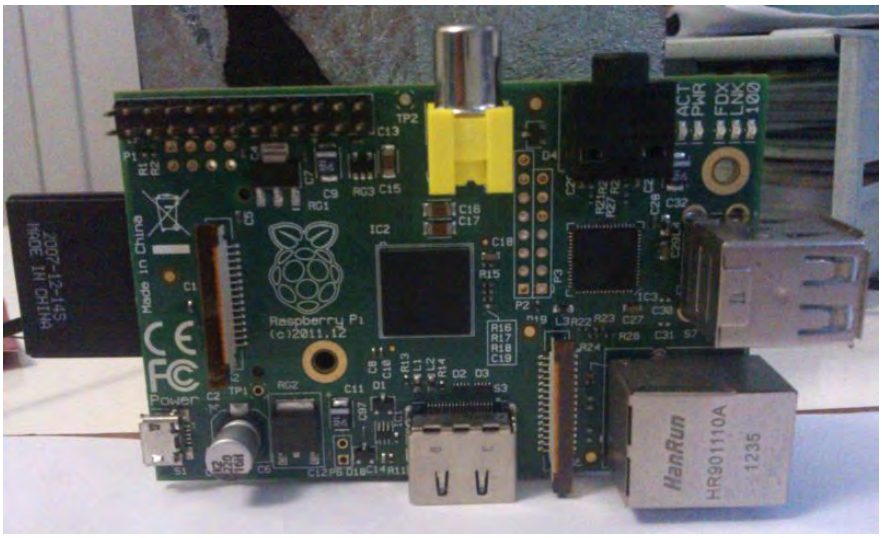
The on-board signal processing of thermal cores incorporates a multi-stage customisable pipeline to condition the information for use. In order to represent thermal scenes, it is recommended that the parameters for one of the many Automatic gain control (AGC) algorithms is tuned for the specific application[4].

The distribution of intensity levels is based on the highest and lowest temperatures in the scene, causing thermal images to often have large areas of low contrast. The resulting image is often mapped onto an 8-bit intensity image (as seen in Fig. 3.3); alternatively, the thermal images are represented in false colour (as seen in Figs. 2.1 and 3.2) to emphasise subtle changes in thermal intensity using 24-bit colour information.

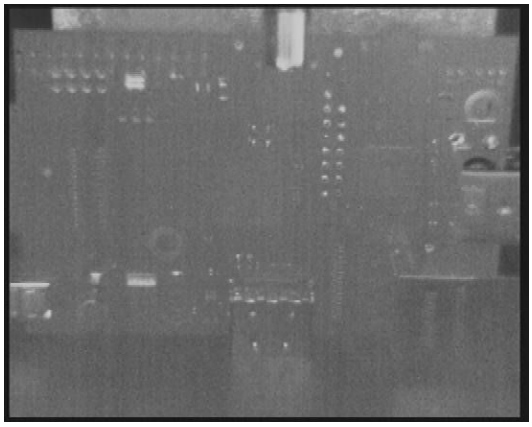
Thermal phenomena are characteristics of thermal images that are not found in the visible spectrum. While the removal of objects in visible scenes is immediately apparent, warm objects leave their thermal impressions on the environment. Phenomena like these, termed 'history effects', can interfere with multi-spectral systems geared towards tracking or change detection. Another characteristic of thermal scenes is what is termed a 'halo effect', which manifests at borders between materials or objects where there is a large temperature difference. This effect, which softens and blurs edges, is due to heat dissipating quickly from a hot to a cold object and emitting thermal radiation as a result.

Working with thermal sensors provides novel challenges in computer vision; however, their operation is largely similar to conventional digital cameras. The next section introduces the multi-spectral camera set-up and the set of methods required for multi-spectral fusion.

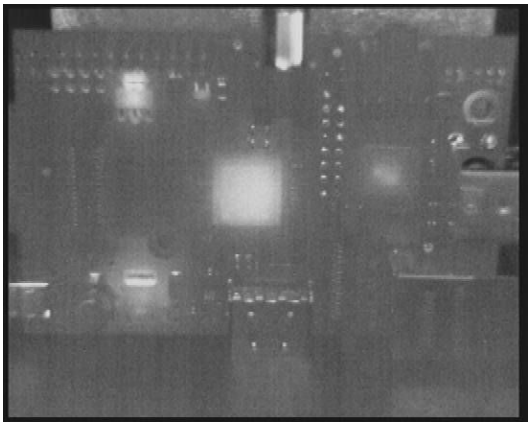
[3]Details available at `www.raspberrypi.org`, August 2015.
[4]FLIR Camera Adjustments and Applications Note, included in camera documentation.



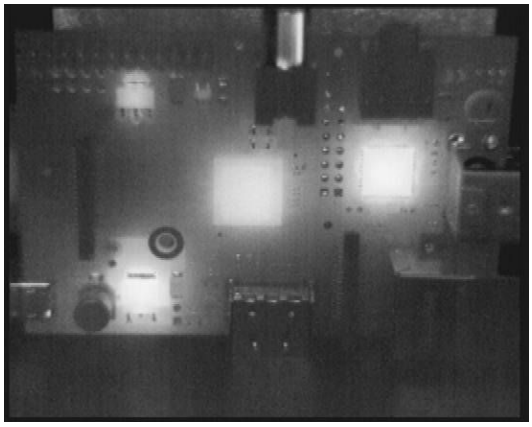
(a) Original visible spectrum image. Component markings, branding and component colours are clearly seen in this image.



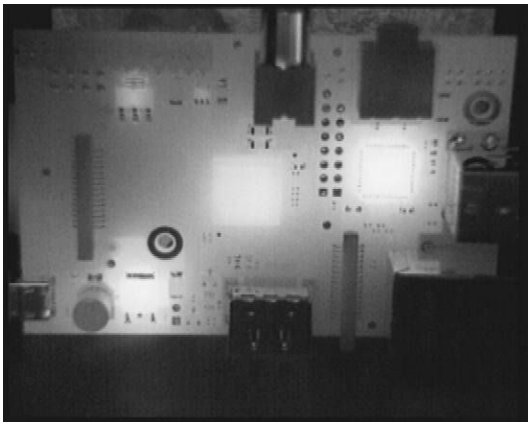
(b) Thermal image at $t = 0$s.



(c) Thermal image at $t \approx 7$s.



(d) Thermal image at $t \approx 14$s.



(e) Thermal image at $t \approx 20$s.

Figure 3.3: A Raspberry Pi mini-computer booting up. The series of images shows the absence of insignia and printed details in the thermal spectrum as well as how the appearance of the thermal image changes based on the temperature range present scene.

3.2 Consolidating multiple views

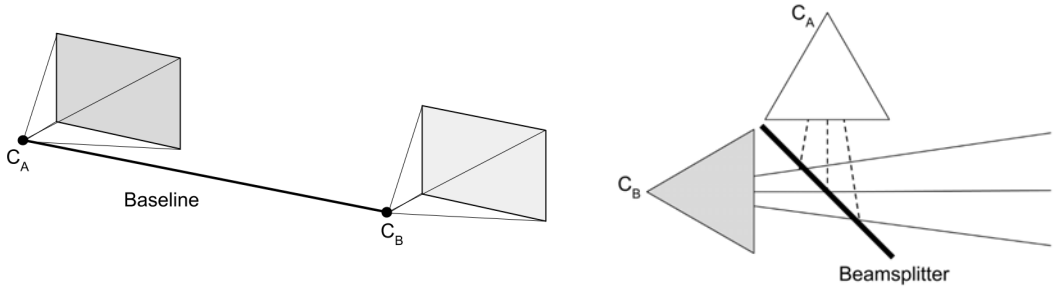
Aerial and satellite applications dominate multi-spectral fusion literature [1, 52]. For these platforms, the distance of the observed scene from the apertures enables the multi-spectral observations to be globally aligned through registration — a pixel-to-pixel mapping that allows the images to be directly overlayed [12]. In a multi-viewpoint (stereo-vision) system, each camera captures a slightly different projection of the scene. If the scene is close enough to the apertures, the observations can no longer be globally aligned.

There are four parts to this section which aim to describe the relationship between the apertures of stereo-vision systems and present the methods that enable us to relate and combine the information captured by these distinct viewpoints. Section 3.2.1 specifies the mounting of the two cameras in this work and how it differs from other approaches. The traditional pinhole camera model is briefly summarised in Section 3.2.2 to illustrate the special considerations required to calibrate thermal cameras. Section 3.2.3 describes the essential purpose of alignment and rectification in fusion. Finally, the decision to use computational stereo methods is presented in Section 3.2.4; the two methods selected for development in this work are specified and the objectives of the investigations are clarified.

3.2.1 Camera configuration

Figure 3.4 presents two camera set-ups commonly used to capture visible and thermal sensor information. The configuration used in this project is shown in Fig. 3.4a. The two front-facing cameras are separated by a horizontal baseline which connects the optical centers of the cameras [53]. The displacement between the apertures introduces parallax. Parallax is the non-linear relationship between the perceived displacement of a point when observed by each aperture and the depth (i.e. distance from the baseline) of the point in the scene [54]. Objects seen from the different viewpoints are therefore related by non-linear displacement (or disparity) related to each object's depth; the result is that the two viewpoints cannot be globally registered and overlayed.

Figure 3.4b shows the use of a beam splitter to create a common aperture in order to remove the baseline between views; it only remains to calibrate the cameras to register the two images with a single planar transform. Beam splitters are expensive (thermal



(a) Two front-facing cameras separated by a horizontal baseline.

(b) A beam splitter is used to create a common aperture to eliminate parallax.

Figure 3.4: Two common dual-camera configurations [53]. The front-facing setup in (a) is used in this dissertation.

7–14 μm zinc selenide beamsplitter prices range from \$695.00 at 25.4mm^2 to \$1,195.00 at 50.8mm^2)[5]and require precision machining of a rigid casing in which to align and mount the two cameras and beam splitter.

The inclusion of parallax in the camera model introduces significant challenges and dictates the approach to fusion; however, to make the problem more tractable, the two viewpoints can be calibrated and brought into alignment. The remainder of the chapter discusses these two processes.

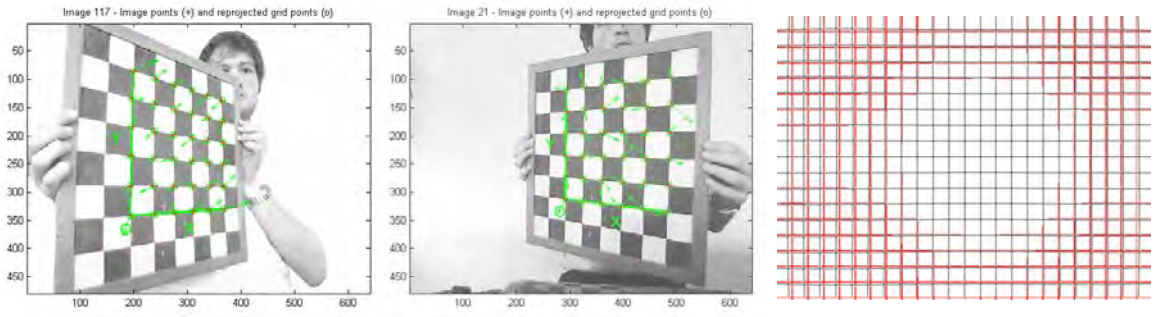
3.2.2 The camera model

The pinhole camera model is widely used to describe a generic camera aperture [48]. Hartley and Zisserman [55] provide a comprehensive mathematical guide to general projective camera models, although a functional description of the pinhole camera model is provided by many sources [48, 56, 57].

When a camera is manufactured, certain intrinsic errors can creep into the various components which cause warping of the image plane. Radial and tangential distortion are caused by non-uniform focal length from the optical centre and misalignment of the lens and sensor respectively. These errors, internal to the camera, are corrected through a process called intrinsic calibration [58]. Tools like the MATLAB Calibration Toolbox [56] and the OpenCV library [57] can be used to estimate the parameters required to perform this calibration.

[5]Plate IR beam splitters priced at `http://www.edmundoptics.com`, August 2015.

Figure 3.5 provides an example of calibration of visible-spectrum cameras using a checkerboard pattern on a planar surface. The inner junctions of neighbouring squares, called saddle points, are used as control points for calculating the parameters required for calibration. An example of an image used in the camera calibration process using MATLAB is shown in Fig. 3.5a. Figure 3.5b shows the calibration transform applied to a regular grid pattern: the modified grid is superimposed onto the original pattern to show how the image is modified by the calibration transform.



(a) Camera calibration sample using a chess board.

(b) Calibration transform.

Figure 3.5: Camera calibration is performed on a single camera. Image (b) shows the how the corrective calibration transform warps the image plane.

Recall that the thermal intensity of an object is a function of its temperature and material emissivity and is independent of colour. Checker-board patterns with modified materials are used for multi-spectral stereo calibration to ensure the sufficient contrast between adjacent squares required to establish these control points. Vidas et al. [51] propose the use of an absorptive mask over a highly emissive surface to produce high contrast control points. The CVC Multimodal Stereo Dataset was rectified using a laser printed checker-board pattern on thin aluminium sheet [35, 36].

3.2.3 Alignment of multiple viewpoints

Alignment of the images captured by each camera to a common viewing plane is an essential calibration step. Stereo calibration is often performed simultaneously with the intrinsic calibration of each camera; the goal is to estimate the extrinsic parameters (i.e. the orientation of the apertures relative to each other) to formulate projective transforms for each image that brings the image planes into partial alignment along rows in the image pair [12]. This process is called rectification.

The relationship between the two viewpoints is not a pixel-to-pixel mapping; instead it is a pixel-to-line mapping, defined by the fundamental matrix. Information provided by the fundamental matrix can be used to rectify the images. Figure 3.6 illustrates the relationship between a point in one image and the line segment, called an epipolar line, along which it can be observed in the other. The rectification process transforms both viewpoints such that these epipolar lines are horizontal and aligned (i.e. features occur on the same row in both images). This relationship is called the epipolar constraint and restricts the search for a corresponding observation of a point in either viewpoint to a 1D search along the corresponding row in the other.

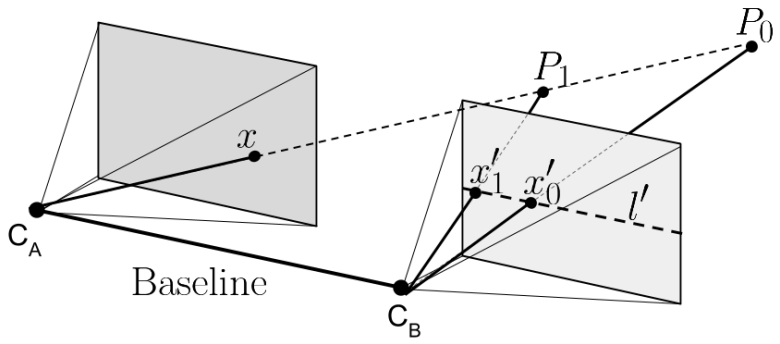


Figure 3.6: An illustration of the fundamental matrix and rectified image planes. The point x on the image plane of C_A could occur at any point along the bold line from C_A through example points P_0 and P_1. The rectified image planes mean that the search for the point corresponding to x observed by C_A is a one dimensional scan along the line l' in C_B.

The output of the stereo calibration process is a rectified image pair. The accuracy of this alignment is crucial to the next step, which is to overlay corresponding regions.

Automatic detection and matching of control points is commonly performed to calculate the rectifying homography of visible spectrum images; automatic rectification by matching elements from the environment is very useful, especially if the camera rig is not accessible or needs to be re-calibrated on a regular basis. However, based on the literature examined at the time of writing, alignment of visible and thermal images is always done manually (or through extensive calibration [23]). To address this, an investigation into the feasibility of automatic rectification using conventional correspondence matching methods is carried out and the results presented in Chapter 5.

3.2.4 Computational stereo

Computational stereo refers to the process of recovering three-dimensional structure of a scene from two or more viewpoints [59]. The process is divided into three stages: calibration and alignment, correspondence search and matching, and reconstruction and disparity refinement. The first stage (calibration and alignment) is discussed in the previous section.

Correspondence search and matching is the next major step and involves finding where a point in the scene is projected in the two image planes. The search for corresponding observations in rectified images is restricted to one dimension along a pixel row; the horizontal displacement between observations is called disparity.

Local methods of computational stereo include block/region matching, gradient methods and feature matching [59, 60]. Gradient methods (e.g. optical flow) rely heavily on brightness constancy across views and therefore cannot be used with visible and thermal images [59]. Local region matching (also called windowing or block matching) was chosen as the method for generating correspondences as disparity selection is a simple process of selecting the point along the search domain that maximises a similarity metric. This method is easily adapted and enhanced, and the components can be generalised and incorporated into existing methods.

The reconstruction stage of the computational stereo process uses the estimated disparity of scene elements (e.g. pixels, features or regions) to generate a disparity map: a dense pixel-to-pixel map representing the estimated disparity for every pixel. At this stage of processing, constraints are used to enforce global (e.g. left-to-right ordering, monoticity) and local (e.g inter-row) consistency. Detecting and compensating for visual occlusion (i.e. scene elements only observable from one viewpoint) is a challenging problem which is compounded by the presence of non-simultaneous phenomena — scene elements such as edges and textured regions that are observable in one spectral modality only. Traditional windowing methods are sensitive to occlusion and tend to not work well in large textureless regions, although significant improvements on the simplistic fixed-size square windowing approach, including multi-scale and spatially adaptive windowing methods, have been made to increase robustness and decrease sensitivity to depth discontinuities and occlusion [60–62].

It has been shown that traditional computational stereo methods fail to identify

correspondences between thermal and visible images in natural scenes [47]. Prior to Barrera et al. [34] there had been no published results of dense disparity estimation with thermal and visible spectrum images. The first two stages of computational stereo (i.e. alignment and correspondence matching) both require a means of comparing regions of the disparate modalities; this common requirement is the focus of Chapter 4 of this work. It should be noted that plane sweep algorithms provide an interesting alternative approach to the problem of computational stereo and fusion, but are not investigated in detail in this report [63]. Chapter 6 presents the development and evaluation of a similarity metric aimed at local region matching for computational stereo.

3.3 Summary

The goal of this chapter is to provide context to the objectives of this work and to clarify the approach to fusion dictated by the camera configuration. Thermal phenomena are discussed in order to introduce the challenges associated with multi-spectral image fusion. The calibration and correspondence matching stages of computational stereo are selected for further study due to the common requirements of the two methods.

Both calibration and correspondence matching require a method to compare the information from the disparate thermal and visible spectra. Presented in the next chapter is the development of a method for extracting stable information shared by the distinct modalities; the goal is to establish a similarity metric that can be used to compare regions of multi-spectral data. An investigation into the performance of conventional and adapted methods for automatic alignment and calibration of multi-spectral viewpoints is presented in Chapter 5, followed by Chapter 6, where a cost function is developed to estimate a mapping function that can be used to overlay the multi-spectral information.

Chapter 4

Phase congruency

The appearance of a scene (regardless of the spectral range) can vary considerably between different observations. This challenge is compounded when the observations capture distinct spectral information, which varies independently in each modality. In order to compare regions from different spectral sensors, a means of extracting scene elements which are repeatable and stable across these variations is required.

This chapter presents a frequency-domain analysis tool called phase congruency. An overview of what is required from phase congruency, its purpose in fusion and the goals of this chapter are presented in Section 4.1. Section 4.2 provides a detailed formulation of phase congruency; each step is discussed in the context of multi-spectral information and the goal of extracting a stable measure to compare the disparate modalities. An investigation which aims to identify predictable quantities extracted in the phase congruency analysis process is specified and carried out. Section 4.3 clarifies the objectives of this investigation and the approach to achieving them. The results are then presented and discussed in Sections 4.4 and 4.5 respectively.

4.1 Introduction

Although thermal images appear very different from the familiar visible spectrum, material boundaries and object edges still make elements in the scene recognisable to the human eye. This observation motivates the use of structural features (e.g. edges,

contours and corners) as stable and matchable elements of multi-spectral scenes. Edges are commonly extracted using the Canny edge detector [64] in which linear filtering with a heuristic threshold is used to detect edges at points of high contrast, i.e. a large gradient magnitude between adjacent pixels is used to indicate the presence of an edge. The undesired effect of this method is that areas of low contrast are left devoid of features and, due to the fixed-size filter, blurred edges are poorly localised [40]. Thermal images in particular often contain large areas of low contrast, and edges are often blurred due to the low spatial resolution and thermal phenomena discussed in Section 3.1.2. Therefore, structural features cannot be repeatably detected or compared using conventional gradient-based methods.

Phase congruency is an extension of the local energy model, which utilises the relationship between structural (or spatial) features in images and their underlying Fourier components [65,66]. This earlier formulation of phase congruency found limited success due to its sensitivity to noise and poor localisation of features [41]; however, a number of works have extended this theoretical base to provide a measure of absolute structural significance particularly well suited to detecting spatial features in natural scenes with large areas of low contrast [40–42, 67]. Its use in multi-spectral imaging is motivated by its independence from image contrast, ability to accurately localise features in images with low spatial resolution and robustness against image noise.

This chapter presents an investigation into the phase congruency process with two objectives. The first is to develop an invariant representation which depicts the stable information shared by both spectral modalities. Creating a visually similar representation allows methods that compare visual similarity between regions to be adapted to function in any spectral range by incorporating phase congruency as a preliminary stage of processing. The second objective is the focus of the investigation and is aimed at identifying predictable relationships between elements exposed in the phase congruency analysis process. Extracting information that can be used to reliably compare the disparate modalities is an important step towards matching corresponding points captured by the distinct sensors.

4.2 Development of theory

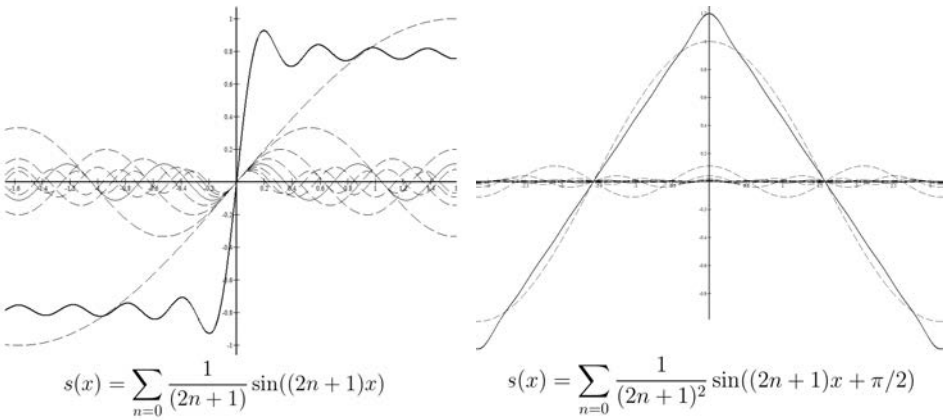
Phase congruency is calculated on images using directional (oriented) log-Gabor multi-scale filter banks in which each point is analysed individually within the context of

its neighbours. An overview of the aims of the phase congruency process is provided in Section 4.2.1. Frequency-domain analysis is performed using a series of multi-scale filter banks; Section 4.2.2 introduces the adapted Gabor filters used to perform this analysis on each pixel. Phase congruency is initially formulated in one dimension in Section 4.2.3, and is then extended to two dimensions for use in image processing in Section 4.2.4. A final stage of computation, presented in Section 4.2.5, condenses the directional congruency values into a reduced feature set which describes the invariant information in the visible and thermal spectra.

4.2.1 Overview

Phase congruency provides invariance to the distinct intensity information of thermal and visible spectrum scenes by performing detection in the frequency domain. The analysis process used to detect spatial features has close parallels to the primary visual cortex of mammals – visual systems that are particularly adept at compensating for changes in contrast, brightness and scale (blur) [68–70].

Edges are perceived at points where the Fourier components are maximally in-phase [40]; Figure 4.1 shows the in-phase Fourier components of two spatial features, an edge and a roof, to demonstrate this. Analysis in the frequency domain provides a means of characterising and detecting features that is adaptive to the scene content and independent of the variations between the spectral modalities [40].



(a) Fourier approximation of a step. (b) Fourier approximation of a roof.

Figure 4.1: The two structures (solid line) are approximated by maximally in-phase (at $x = 0$) the Fourier components (various dotted lines) [41, 42].

4.2.2 Gabor filter banks

A *feature* (in this context) is a point of phase congruency where the Fourier components are maximally in phase over a large range of frequencies. The first stage of processing is therefore to extract a large number of frequencies so that the signals can be analysed for congruency. Uniform coverage of a broad range of frequencies is extracted using a multi-scale bank of band-pass filters [41].

Gabor filters are band-pass filters that allow us to isolate frequency components of a signal. They are commonly found in contour extraction, texture analysis and object recognition applications (a taxonomy presented in [71]). In this work, they are used for calculating localised frequency content in a signal [39].

The impulse response of a Gabor filter is a Gaussian distribution (kernel) with standard deviation σ, modulated by a complex sinusoid (carrier) at frequency ω_0 [72]; the frequency domain representation is plotted in Fig. 4.2.

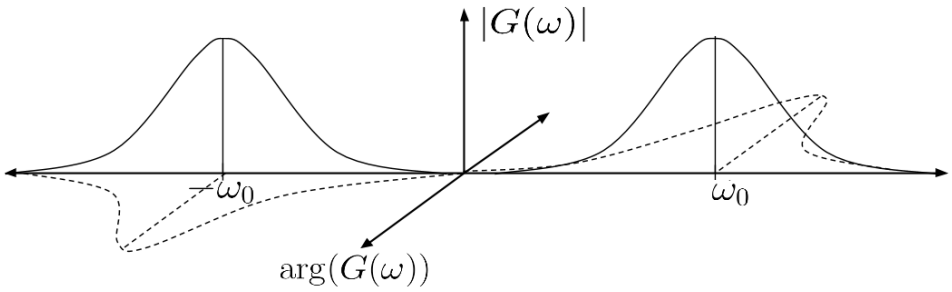


Figure 4.2: Transfer function of the Gabor band-pass filter operating at frequency ω_0.

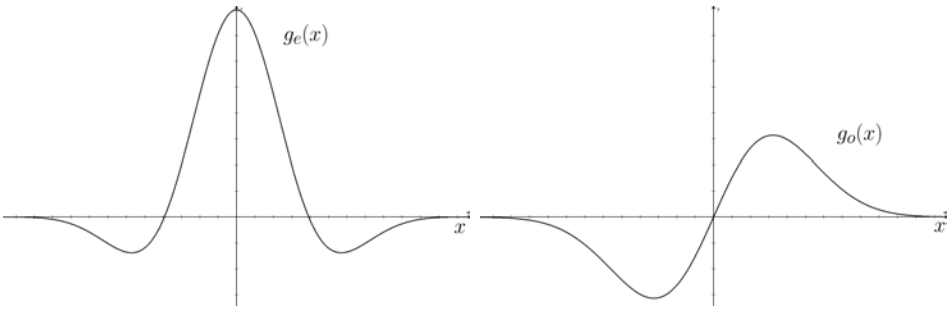
Gabor filters are represented as a quadrature pair of orthogonal (sinusoids offset by $\pi/2$), real-valued even and odd symmetric filters:

$$g(x) = g_e(x) + jg_o(x). \tag{4.1}$$

The spatial profile of each component is expressed as

$$g_e(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x^2}{2\sigma^2}} \cos(2\pi\omega_0 x) \quad \text{and} \tag{4.2a}$$

$$g_o(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x^2}{2\sigma^2}} \sin(2\pi\omega_0 x). \tag{4.2b}$$



(a) Even-symmetric Gabor profile. (b) Odd-symmetric Gabor profile.

Figure 4.3: 1D spatial profiles of a Gabor filter quadrature pair.

The spatial profiles of these Gabor filters are shown in Fig. 4.3.

The Gaussian envelope is approximately zero at three standard deviations from the mean of the envelope (at the carrier frequency ω_0). A DC component, illustrated in Fig. 4.4, is induced if the carrier frequency is too low or if the envelope σ is too admitting. Consequently, a bandwidth limitation on the Gabor filter constrains its coverage of the spectrum.

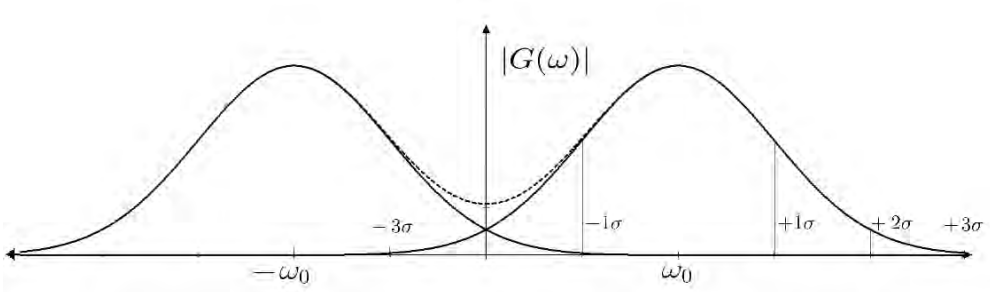


Figure 4.4: A significant DC component is introduced if the center frequency is smaller than 3σ of the Gaussian envelope.

The effect of this limitation, as presented in [39], is that many more band-pass filters are required to obtain adequate coverage of the spectrum. A modified form of the Gabor filter is therefore used for the phase congruency process to overcome this limitation.

Log-Gabor filters

Log-Gabor filters allow us to specify an arbitrarily large bandwidth without inducing a DC component in the even-symmetric filter [39, 41, 73]. The result is that the bandwidth limitation is lifted and the computational cost of analysing a broad range

of frequencies is decreased (as fewer filter scales are required). A filter with a smaller foot-print provides better spatial localisation at the cost of coverage of the spectrum; however, a log-Gabor filter is able to cover three octaves, as opposed to a regular Gabor filter of the same width which only covers one [39]. Therefore, log-Gabor filter banks provide better spatial localisation in addition to the broad frequency coverage.

Field's study of the statistics of images found that natural scenes have a spectral profile that falls off as roughly $1/\omega$ [73]. Gabor filters, due to their Gaussian envelope, tend to over-represent lower frequencies [39]; however, the similar spectral profile of natural scenes and the log-Gabor filter response, Field argues, makes log-Gabor better suited to encoding the spectral information of natural scenes.

Coverage of the spectrum is achieved by geometrically scaling the wavelength of the log-Gabor carrier signal. The next section describes how the multi-scale log-Gabor filter bank is constructed.

Scaling bandpass filters

The multi-scale filter bank used in the phase congruency process consists of *nscale* log-Gabor filters geometrically scaled from a minimum wavelength λ_{min} with a constant wavelength multiplier λ_{mult} [41]. The center frequency ω of each successive filter at scale n is

$$\omega_n = \frac{1}{\lambda_{min}\lambda_{mult}^{n-1}}. \tag{4.3}$$

The minimum frequency (maximum wavelength) is indirectly determined by the specification of the minimum wavelength, number of scales and scaling factor parameters. As there is no analytical expression for the log-Gabor envelope, due to the singularity at the origin, it is specified in the frequency domain as [39]

$$G(\omega) = e^{\frac{-(\log(\omega/\omega_0))^2}{2(\log(\kappa/\omega_0))^2}}. \tag{4.4}$$

The κ/ω_n factor is kept constant for varying ω_n to ensure a constant shape ratio for uniform coverage of the spectrum [41].

Phase congruency is a process that infers structural features from the frequency information extracted by the multi-scale log-Gabor filter bank presented here. The

next sections introduce the phase congruency method in one dimension using the tools developed so far; these methods are then extended to 2D for use with images.

4.2.3 Multi-scale congruency on a one dimensional signal

Phase congruency assigns an absolute significance to points where the Fourier components are maximally in phase [41]; this implies that frequency components exist over a broad range of the spectrum and that there is consensus of local phase between the components. The following sections formulate the phase congruency measure.

Filter response

At each point in the signal, the log-Gabor filter bank generates a set of even and odd response vectors at each scale. Components of the response vector, $e_n(x)$ and $o_n(x)$, are formed by convolving the quadrature pairs of even and odd filters at each scale n with the underlying signal information. The filter responses are described in terms of the amplitude, A_n, and phase, ϕ_n, at each scale defined by

$$A_n(x) = \sqrt{e_n(x)^2 + o_n(x)^2} \quad \text{and} \tag{4.5a}$$

$$\phi_n(x) = \tan^{-1} \frac{o_n(x)}{e_n(x)}. \tag{4.5b}$$

The magnitude of the response is used in calculating frequency spread, and the phase of response is used to quantify consensus on the underlying structure over multiple scales.

Weighting for frequency spread

Frequency spread is based on the relative filter response magnitudes at each scale. The output, $W(x)$, is a scalar weighting function indicating the presence of Fourier components over a broad range of frequencies.

Frequency spread, $s(x)$, is a normalised accumulation of the response magnitudes at

each scale relative to the strongest response A_{max} at that point:

$$s(x) = \frac{1}{N}\left(\frac{\Sigma_n A_n(x)}{A_{max}(x) + \epsilon}\right). \tag{4.6}$$

A small constant, ϵ, is included in the denominator to avoid division by 0. Frequency spread is incorporated into a weighting function which is used to penalise points where there is a narrow range of frequencies:

$$W(x) = (1 + e^{\gamma(c - s(x))})^{-1}. \tag{4.7}$$

Parameters c and γ control the cut-off (below which the phase congruency value is penalised) and the steepness of the curve respectively. The weighting function is plotted in Fig. 4.5.

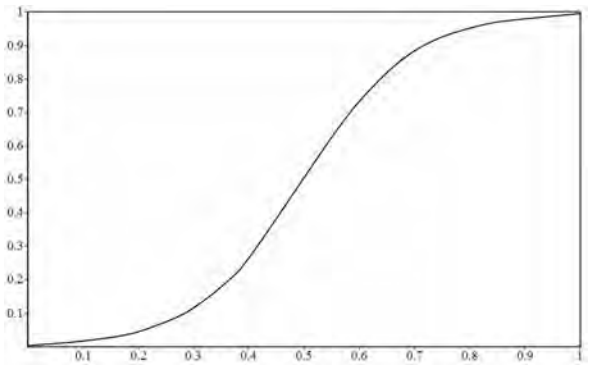


Figure 4.5: Multi-scale weighting function used to calculate frequency spread from filter response magnitudes [41]. Values below the cutoff parameter, c, are penalised. (In this case $c = .5$ and $\gamma = 10$.)

Note that this measure is based on the relative weighting of filter responses and is independent of the intensity information (i.e. gradient magnitude or brightness) of the image.

Quantifying consensus on underlying structure

Local phase is based on the ratio of quadrature filter responses at each scale, and is often depicted, described and compared as an angular quantity (shown in Eq. (4.5b)). Phase is an indication of the underlying structure at a point in the signal. As congruency occurs at points where the Fourier components are maximally in phase, a means of quantifying the consensus of these components is required.

The mean phase angle $\bar{\Phi}$ is a normalised direction vector representing the combined phase consensus over multiple scales and is defined as:

$$\bar{\Phi} = \frac{1}{\sqrt{(\Sigma_n e_n)^2 + (\Sigma_n o_n)^2}} (\Sigma_n e_n, \Sigma_n o_n). \tag{4.8}$$

Phase consensus is quantified by calculating the average (mean) phase angle and then weighting each filter response magnitude A_n with the vector's deviation from this mean phase angle.

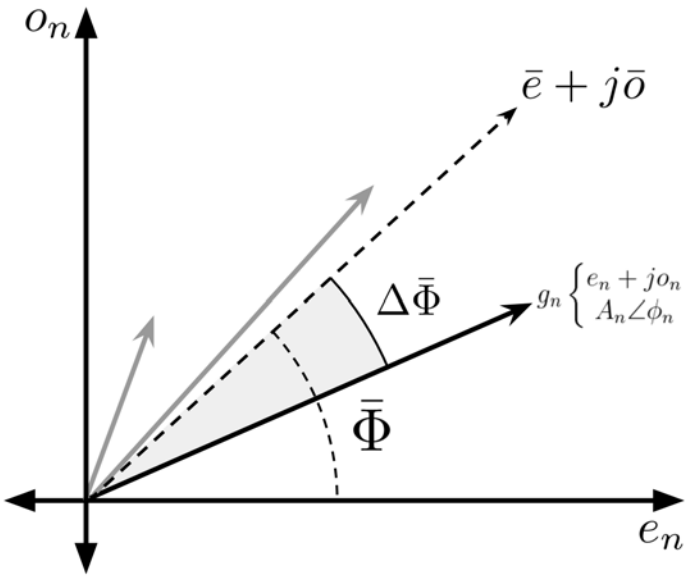


Figure 4.6: The mean phase angle $\bar{\Phi}$ is calculated using the filter responses at each scale. An arbitrary filter response g_n with its equivalent representations is shown, and its phase deviation $\Delta\bar{\Phi}$ is labelled.

Figure 4.6 illustrates the phase deviation $\Delta\Phi_n$ of an arbitrary response vector g_n. The phase deviation weighting function is based the angular distance between the phase angles and is formulated as

$$\Delta\Phi_n = \frac{1}{2}(\cos(\phi_n - \bar{\Phi}) - |\sin(\phi_n - \bar{\Phi})| + 1) \tag{4.9}$$

and plotted in Fig. 4.7.

As the magnitude and phase deviation is to be calculated at each point and at each scale in the image, the calculation is simplified to remove trigonometric equations and

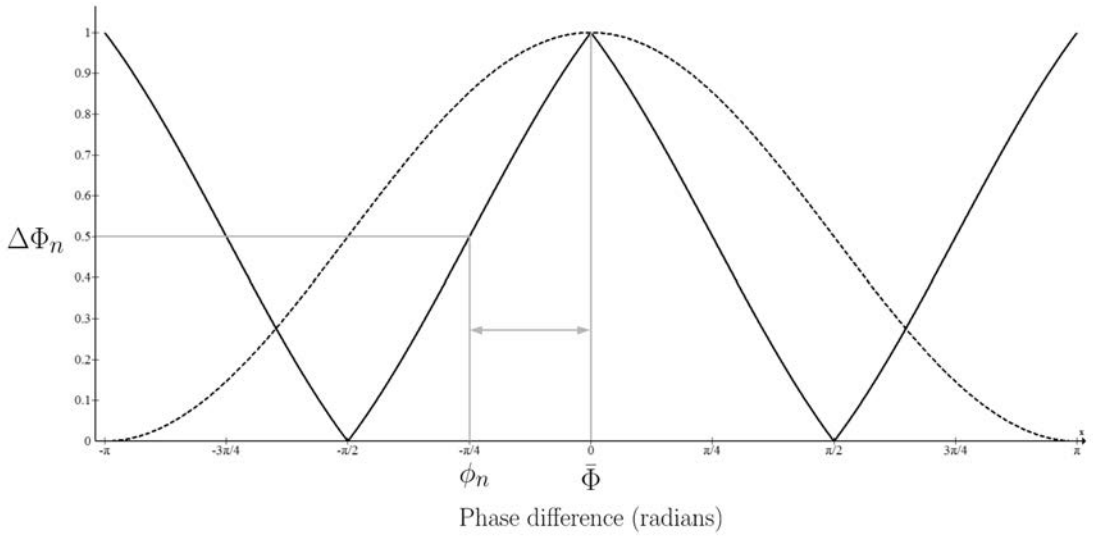


Figure 4.7: A comparison of the cosine distance function (dashed line) and the phase deviation measure (solid line). The phase deviation weighting $\Delta\Phi_n$ is calculated using the difference of the phase response ϕ_n and the mean phase angle $\bar{\Phi}$ over all scales.

costly divisions. The expansion and modified function are stated as [40]

$$A_n\Delta\Phi_n = \sqrt{e_n^2 + o_n^2}\ \big(\cos(\phi_n - \bar{\Phi}) - |\sin(\phi_n - \bar{\Phi})|\big) \tag{4.10a}$$

$$A_n\Delta\Phi_n = e_n\bar{e} + o_n\bar{o} - |o_n\bar{e} - e_n\bar{o}|. \tag{4.10b}$$

Each filter response contributes $A_n\Delta\Phi_n$ — magnitude, weighted by its consensus with the mean phase angle — to the phase congruency measure. However, as natural image contain noise, the response of each filter needs to be significant enough to be distinguishable from noise.

Compensating for noise in the signal

Normalisation is required in any adaptive measure, although this induces a sensitivity to noise [41]. Image noise is assumed to be additive and features (edges and corners) are assumed to be sparse and isolated so that they may be distinguished from background noise. In order to quantify the expected noise level within the image, the response magnitude of the filter with the smallest spatial extent (making it the most sensitive to noise) is used to establish a noise floor.

The magnitude of the noise follows a Rayleigh distribution [41]. The noise threshold

T can be tuned in terms of standard deviations from the mean of this distribution.

Calculating phase congruency

The phase congruency measure PC is calculated at each point x in the signal as

$$PC(x) = \frac{W(x)\Sigma_n \lfloor A_n(x)\Delta\Phi_n(x) - T \rfloor}{\Sigma_n A_n(x) + \epsilon}. \tag{4.11}$$

At each scale n, the filter response magnitude A_n is weighted for phase deviation $\Delta\Phi_n$ and reduced by the estimated noise floor T. A thresholding operation $\lfloor A_n(x)\Delta\Phi_n(x) - T \rfloor$ is applied at each scale of filter response. The $\lfloor \cdots \rfloor$ notation indicates that if the enclosed result is negative (the filter response lying below the noise floor) then the result is equal to 0 else it is left unchanged. The summation in the numerator then is normalised using the filter response magnitudes in the denominator (ϵ is a very small constant to avoid division by zero). The entire calculation is then weighted for frequency spread $W(x)$.

The result of the phase congruency measure, presented in Eq. (4.11), is a value between 0 and 1 that indicates the significance of the underlying structure (degree of congruency) of a one dimensional signal.

4.2.4 Congruency over multiple orientations

Extending the phase congruency theory into two dimensions requires a modification of the log-Gabor filter bank to include orientation selectivity. The aim is to apply the one dimensional theory in discrete orientations and then combine the oriented results into a descriptive quantity for each pixel.

Directional filter bank design

Two dimensional (spatial) Gabor filters are used to extract local frequency content (band-pass) in discrete directions [70]. Gabor filters are extended to two dimensions

as follows:

$$g_e(x) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x}+\frac{y^2}{\sigma_y}\right)} \cos(2\pi\omega_{x_0}x + 2\pi\omega_{y_0}y) \tag{4.12a}$$

$$g_o(x) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x}+\frac{y^2}{\sigma_y}\right)} \sin(2\pi\omega_{x_0}x + 2\pi\omega_{y_0}y). \tag{4.12b}$$

Figure 4.8 shows a contour plot of the spatial log-Gabor filter over multiple scales; it can be seen that log-Gabor filters have a logarithmic Gaussian profile in the radial direction (outwards from center) and a standard Gaussian profile in the angular direction.

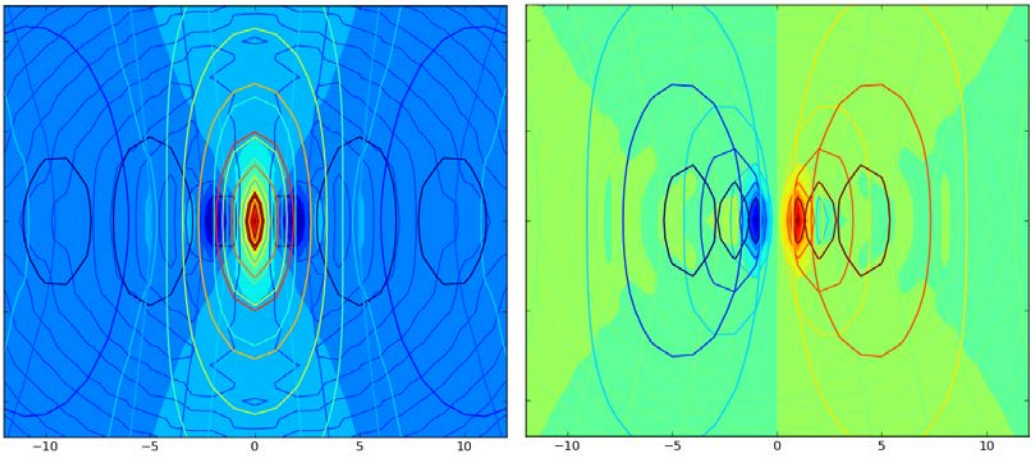
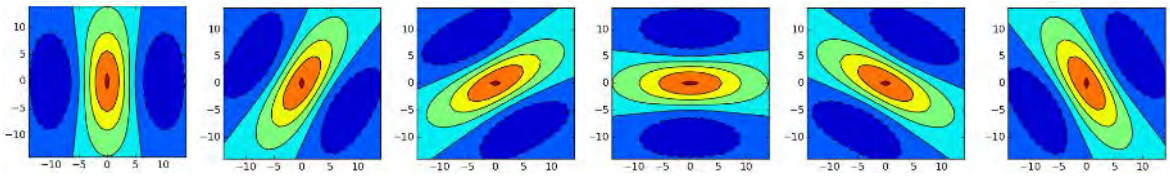


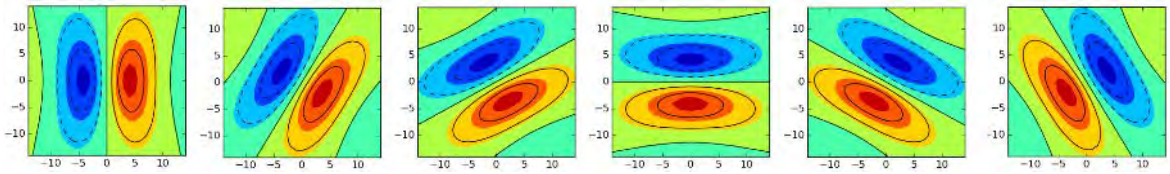
Figure 4.8: A false-colour contour plot showing the superimposed spatial filter responses at four scales (key: dark red is large and positive, dark blue is large and negative, unit is pixels).

By default, six discrete orientations are used which evenly divides the 0 to π range as shown in Fig. 4.9. The number of oriented filters determines the orientation selectivity; however, increasing the number of orientations significantly increases computation time with little added value [67]. Figure 4.10 shows the log-Gabor frequency response contours for 6 orientations at 4 scales.

Filter bank design is a trade-off between *uniform* and *efficient* coverage of the spectrum. In order to obtain uniform coverage, every point of the range of frequencies we want to capture must be uniformly weighted when the filter responses are summed; this requires frequency-adjacent filters to overlap (as illustrated in Fig. 4.10). However, this overlap reduces the efficiency of our encoding as the filters become less independent



(a) Even oriented filter set.



(b) Odd oriented filter set.

Figure 4.9: Even and odd spatial filter responses at six orientations at a single scale (key: dark red is large and positive, dark blue is large and negative, unit is pixels).

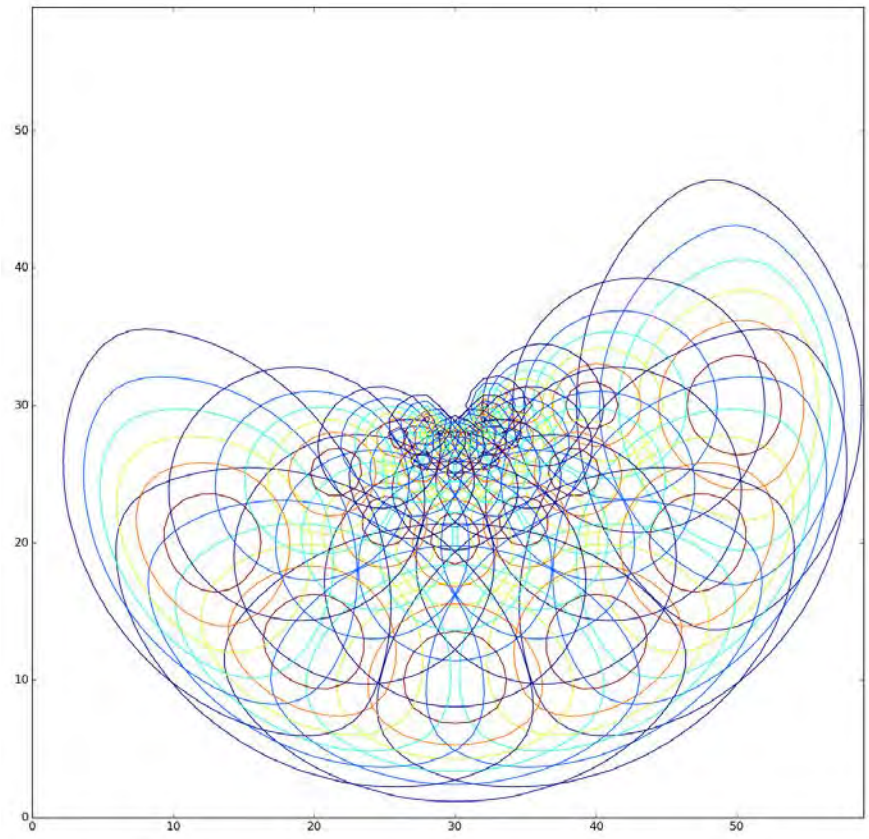


Figure 4.10: A contour plot showing one-sided superimposed oriented spatial frequency responses of the log-Gabor filter bank in the Fourier domain. Note the angular and radial overlap that forms a part of the design trade-offs towards uniform spectrum coverage.

and more correlated [39]. The extension to two dimensions necessitates the inclusion of sigma parameters to control radial and angular spread. The parameters suggested by Kovesi are provided in Section 4.3.4.

As an alternative to the discrete directional filters used in Kovesi's phase congruency, Mellor et al. [74] use steerable band-pass filtering (a difference of Gaussians operation followed by the Hilbert transform) to recover the symmetric phase information. The non-linear log-Gabor filters used in Kovesi's phase congruency are not steerable, which is why calculation in discrete orientations is required.

Oriented phase congruency

Phase congruency is calculated at each orientation using

$$PC(x)_i = \frac{\Sigma_n W(x) \lfloor A_n(x) \Delta\Phi_n(x) - T \rfloor}{\Sigma_n A_n(x) + \epsilon}. \tag{4.13}$$

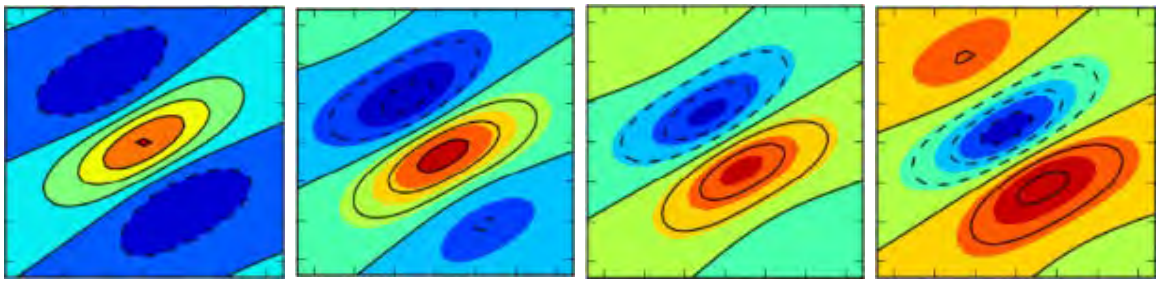
This equation is essentially unmodified from the one dimensional analysis, although oriented components have been appended with a subscript i to indicate that the calculation is specific to each orientation. Note that the noise floor T and the mean phase angle $\bar{\Phi}$ quantities are calculated for each orientation. For the remainder of this report, oriented components are subscripted with i; most notably, the phase congruency PC_i and the mean phase angle $\bar{\Phi}_i$ quantities are collectively referred to as the oriented components of phase congruency.

Figure 4.11 demonstrates the relationship between the phase (i.e. the ratio of the quadrature filter responses) and the underlying structure. The four images show different phase responses of a $\theta = \pi/2$ oriented filter which have been created using

$$P(x, y) = \alpha e_{n,i}(x, y) + \beta o_{n,i}(x, y) \tag{4.14}$$

where each pixel in the image P is a sum of the spatial filters weighted with $\alpha + \beta = 1$. In the phase congruency equation, the mean phase angle represents the consensus of the structure recovered by a particular oriented filter over all scales.

Another reason for the inclusion of Fig. 4.11 is to emphasise the important distinction between the angle representing the mean phase response $\bar{\Phi}_i$, and the filter orientation θ_i. The magnitude of congruency PC_i is associated with the direction of the filter,



(a) $\phi = 0$. (b) $\phi = \pi/2$. (c) $\phi = \pi$. (d) $\phi = 3\pi/2$.

Figure 4.11: A sequence of contour plots of the spatial log-Gabor filter response, illustrating the relationship between phase ϕ and structure; each representation of the oriented ($\theta = \pi/2$) filter is constructed using a different ratio of the quadrature pair. (Key: dark red is large and positive, dark blue is large and negative, unit is pixels.)

and is represented as a vector at angle θ_i. Visualising the oriented components in this way is useful in the next section, which presents a method for combining the values to describe structural features in images.

4.2.5 Representation using a reduced feature set

The oriented congruency PC_i components, calculated in Eq. (4.13), are synthesised to three values per pixel in a moment analysis method presented in [40]. The three values of this reduced feature set are the perpendicular maximum and minimum moment values represented as the scalars M and m respectively, and the angle of the maximum moment called the principal axis θ'.

The scalar moment values are calculated using the magnitude of congruency associated with each filter orientation:

$$M = \frac{1}{2}(c + a + \sqrt{b^2 + (a - c)^2}) \tag{4.15a}$$

$$m = \frac{1}{2}(c + a - \sqrt{b^2 + (a - c)^2}). \tag{4.15b}$$

The components a, b and c represent summations of the phase congruency magnitudes PC_i calculated with Eq.(4.11) at each orientation θ_i, and are calculated as

$$a = \Sigma_i(PC_i \cos(\theta_i))^2 \tag{4.16a}$$

$$b = 2\Sigma_i(PC_i \cos(\theta_i))(PC_i \sin(\theta_i)) \tag{4.16b}$$

$$c = \Sigma_i(PC_i \sin(\theta_i))^2. \tag{4.16c}$$

The maximum moment indicates feature (i.e. edge) significance and is used to create an edge map of the image [40]. If both moment values are large, the point is classified as a corner; consequently, the corner features present in the image are a subset of the edge map. The angle of the principal axis indicates the orientation of the edge and is calculated as follows:

$$\theta' = \frac{1}{2}\text{atan}\left(\frac{b}{a-c}\right). \tag{4.17}$$

Together, the three components of the reduced feature set at each pixel describe the structural make-up of a scene. These components are an integral part of the methods developed in the remaining chapters of this work.

Figure 4.12 has been included to show how the oriented congruency components, PC_i, are described by the reduced feature set. On the right of Fig. 4.12, the maximum moment with scalar magnitude M and direction θ' values can be seen to describe the significance and orientation of the curved edge; the minimum moment, m, describes the 'cornerness' due to the curve in the edge. Of the six orientations used in analysis, only three were non-zero. The non-zero PC_i magnitudes are shown as vectors on the curved edge (i.e. PC_1, PC_2 and PC_3), and the phase response of each is represented as contour plots on the left (i.e. θ_1, θ_2 and θ_3).

The remaining sections of this report present an analysis of phase congruency as a tool for exposing invariant structure and predictable features in multi-spectral images.

4.3 Identifying predictable image elements

The primary objective of the investigation carried out in this chapter is to extract repeatable and predictable structural elements using the phase congruency process. Supervised learning is used to identify predictable relationships between components extracted from corresponding observations in the two spectra.

Section 4.3.1 provides a brief overview of the support vector machine model for supervised learning. Methods to characterise the performance of the learning model are then introduced in Section 4.3.2 to quantify predictability of image elements. Section 4.3.3 describes the approach undertaken to investigate predictability in multi-

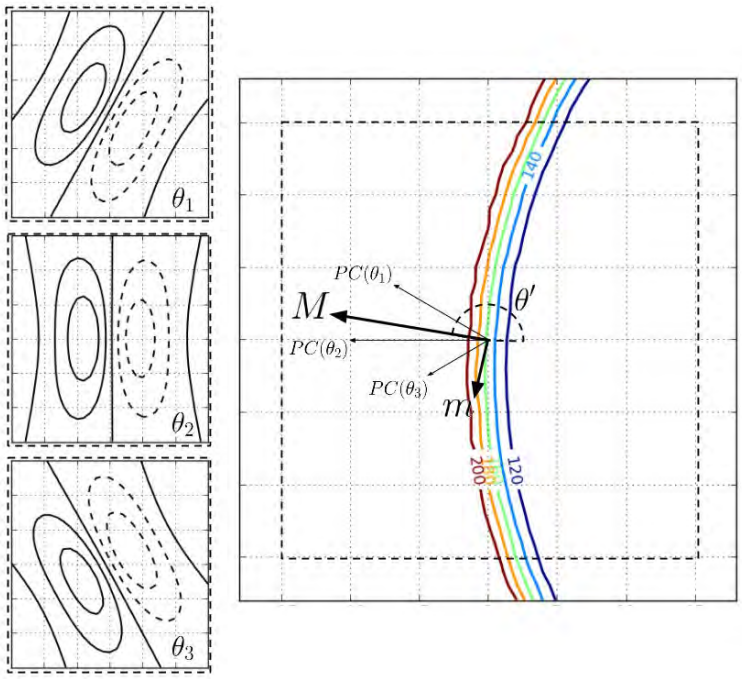


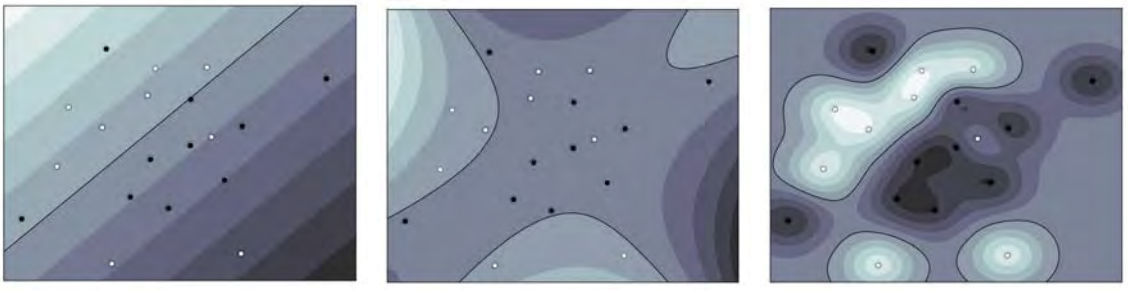
Figure 4.12: A visualisation of the descriptive of the minimum moment m role in quantifying 'cornerness'. The diagram shows the principal axis, described by the angle θ', and the two scalar values of the maximum moment M, colinear to the principal axis, and the magnitude of the minimum moment m, perpendicular to the principal axis. The filters with the maximal responses to the synthetically created curve are shown in the left.

spectral information, and the tools and parameters used in the investigation are recorded in Section 4.3.4.

4.3.1 Support vector machines

Support Vector Machine (SVM) are a class of supervised learning models used in a broad array of applications. They are regarded as one of the best off-the-shelf algorithms for supervised learning and are widely used in multi-spectral and remote-sensing literature [52, 75]. A strong motivation for the use of SVM algorithms in this context is the high accuracy that is achieved with a limited amount of training data [76].

Supervised learning is an iterative process. Sample data with known classes (or labels) is used to train the SVM model which fits a hyperplane that separates the classes. This hyperplane is called the decision boundary, and is used to classify future (unlabelled) samples. Due to the strict requirement of linear separability of classes, a kernel mapping is commonly applied to accommodate data that requires a more complex decision boundary [52]. Figure 4.13 shows a visualisation of three different SVM kernels applied to the same training data on a two dimensional plane. Each training sample is shown as a circle on the plane with a colour indicating its class (black or white). These samples are used to generate a decision boundary, shown as a solid black line, separating the two classes; the shaded levels of grey indicate the confidence with which the SVM classifies each coordinate in the plane (from black to white).



(a) Linear kernel. (b) Polynomial kernel. (c) Radial Basis Function (RBF) kernel.

Figure 4.13: Three SVM kernel variations used to extend the functionality of linear SVM classification to non-linearly-separable data distributions.

The three SVM kernels shown in Fig. 4.13 generate different decision boundaries, which represent different hypotheses that an arbitrary point on the plane will belong to a certain class. In this way, SVMs can be used to discover predictable behaviour and relationships in high dimensional data by formulating and testing different hypotheses.

4.3.2 Characterising performance

The approach to achieving the objectives of this chapter is to use supervised learning with SVMs to identify predictable structure in visible and thermal images; this section describes how the hypotheses are evaluated and what the results can tell us. Three different SVM kernels are used to discover simple (linearly separable) to complex (polynomial and RBF kernels) relationships between components of the phase congruency process.

The trained SVM is required to classify whether two points taken from the different spectral modalities correspond to the same observation in the scene. This is a classification problem to determine if a single pattern vector (a string of numbers composed of values taken from the two points) is indicative of a match or not. To accomplish this, the model requires a set of positive and negative training examples to establish the decision boundary. A training example is a pattern vector with a known class based on the known relationship between the points (i.e. whether the two points used to construct the pattern vector are corresponding observations of the same point or not).

The process of evaluating the hypothesis proposed by the SVM kernel is called testing. A portion of positive and negative samples are kept aside during the training process for use in hypothesis testing. SVM performance is characterised by comparing the classification assigned by the hypothesis to the known class of each sample. Samples with correctly-assigned labels are referred to as true positives and true negatives, while errors in classification are referred to as false positives (incorrectly matched) and false negatives (incorrectly rejected).

Associated with each classification is a numerical value indicating the probability of class membership. By continuously varying a threshold on this probability, a Receiver Operating Characteristic (ROC) plot can be constructed to characterise the effectiveness of the SVM hypothesis. The number of True positive (TP), True negative

(TN), False positive (FP) and False negative (FN) element counts are combined to form the True positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{4.18}$$

and the False positive rate (FPR)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FN}}. \tag{4.19}$$

These two metrics for performance are the axes for the ROC plot. Figure 4.14 shows an example ROC plot in the style presented in this report. The title at the top of the plot shows the components in the pattern vector as well as the length of the pattern vector. The number of training (Tr) and testing (Te) samples used is also shown in the title. Plots of two different kernels are shown: the diagonal line of *kernel1* demonstrates the ROC curve of a useless predictor no better than random guessing, whereas the curve of *kernel2* towards the top left indicates a useful predictor. Comparisons of ROC curves is often based on the Area under curve (AUC) metric, shown in the legend next to each kernel.

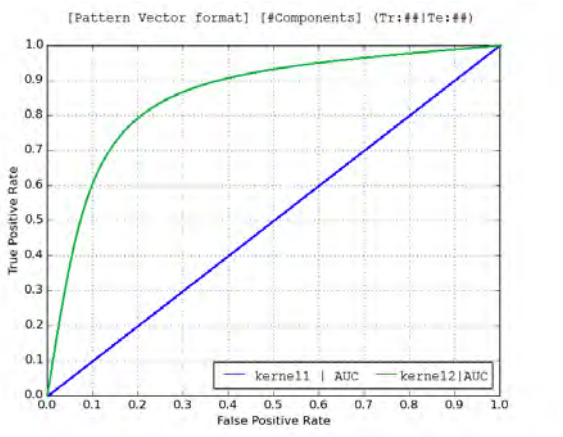


Figure 4.14: An example ROC plot showing two curves. The diagonal *kernel1* shows a classifier that is no better than random binary guessing and *kernel2*, which curves towards the upper left corner, shows a well-performing classifier.

In order to use the ROC plot to characterise and evaluate SVM performance, reliable training and testing data is necessary. The next section describes how the positive and negative sample sets are created from multi-spectral image pairs.

4.3.3 Training and testing data

Samples are extracted from two image pairs using the magnitude of the maximum moment to detect congruent edge features in the images. The training and testing

data was manually separated to ensure that there was no bias introduced into the test by clustering of features. Figure 4.15 shows the partitioned training and testing image regions from the OSU (OCTBVS) Color-Thermal Dataset. In order to avoid the edge effects of filters at the image boundaries, a margin of 15 pixels from the edges of the image is enforced; the training and testing sample data is taken from points in the image that fall inside this border. The size of the margin is chosen to be larger than the envelope of the largest log-Gabor filter used in the phase congruency process, the parameters of which are documented in Table 4.1. The 15 pixel margin used in this investigation was determined in this way, based on analysis of the log-Gabor filter profiles using Octave (software).

Although a number of different pattern vectors (i.e. consisting of different components) were investigated, the method of extracting the positive and negative sample sets was the same. Positive samples were constructed if the corresponding points in both the thermal and visible modalities were edge features. An equal number of positive and negative examples were used in training and in testing; therefore for every positive example, a negative example was constructed by concatenating two randomly selected congruent points from different modalities and image pairs. Thirty percent of the positive examples in the training image pair was randomly sampled to be used to train the SVM with each pattern vector configuration.

The phase congruency process uses both angular (e.g. angle of the principal axis) and scalar (e.g. maximum and minimum moments) quantities. We can take advantage of this extra knowledge to better formulate the pattern vectors by combining and comparing the components accordingly. Angular components can be compared with an adapted cosine distance measure:

$$D(\theta_1, \theta_2) = \frac{1}{2} \left(|\cos(\theta_1 - \theta_2)| - |\sin(\theta_1 - \theta_2)| + 1 \right). \tag{4.20}$$

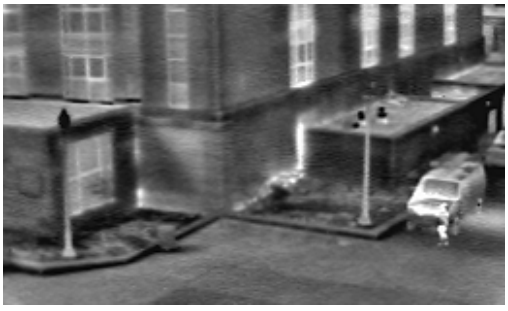
This distance measure compensates for contrast reversals and provides a linear fall-off as the difference between the input angles increases. Scalar components of the phase congruency process are often indicators of significant structure and are paired with an angular quantity. For example, the angle of the principal axis is only meaningful if it is associated with an edge, signified by the maximum moment. Significance measures can be combined by multiplication to indicate that the information being compared is meaningful and not a result of noise. This will become clearer when the results are presented.

(a) Training (pair **A**).

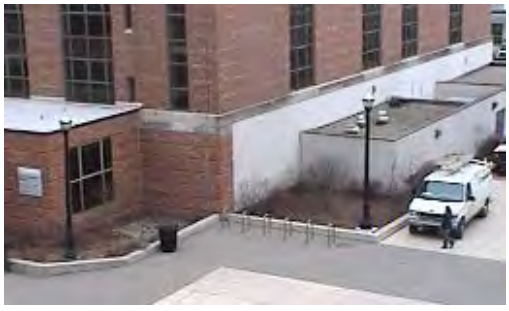
(b) Testing (pair **B**).

(c) Training (pair **A**).

(d) Testing (pair **B**).



(e) Testing (pair **C**).



(f) Testing (pair **C**).

Figure 4.15: Manual segmentation of images into training and testing sections. Training points are randomly selected from edges in training pair **A**. Testing samples are randomly sampled from edges in image pairs **B**, **C**. (Images from the OSU database [44].)

4.3.4 Tools

The phase congruency code used in this project is the MATLAB/Octave implementation provided by Kovesi [77]. The code-base is predominantly Python[1] using the numpy [78], scipy[2] (sklearnkit), matplotlib [79] and OpenCV[3] libraries. The Octave implementation of phase congruency is called from Python using the oct2py[4] module. The default parameters used in the phase congruency process are shown in Table 4.1 [41, 77].

Table 4.1: The default values used in the phase congruency process.

Parameter	Value	Description
nscale	4	Number of filter scales.
norient	6	Number of orientations.
$\lambda_{\min}$	3	Minimum filter wavelength (in pixels).
λ_{mult}	$\{1.3, 1.6, 2.1, 3.0\}$	Multiplicative factor of successive filter wavelengths.
σ_f	$\{.85, .75, .65, .55\}$	Radial sigma value controlling frequency overlap/even coverage (mapped from $\lambda_{\min}$).
k	5	Number of standard deviations from the mean of the noise distribution.
c	0.5	Frequency spread cutoff value (see Fig. 4.5).
γ	10	The sharpness of the sigmoid function.

Note the relationship between the wavelength multiplier and the radial sigma value in controlling radial coverage of frequencies; the default used in this project is $\lambda_{mult} = 1.6$ and $\sigma_f = 0.75$, resulting in a filter bandwidth of one octave with even coverage.

4.4 Results

The objectives stated at the beginning of this chapter were two-fold: to establish an invariant representation of repeatable features of visible and thermal information, and to identify predictable relationships between values extracted in the phase congruency analysis process. The results of these two investigations are presented in this section;

[1] Available at `http://www.python.org`, August 2015.
[2] Available at `http://www.scipy.org`, August 2015.
[3] Available at `http://www.opencv.org`, August 2015.
[4] Available at `http://pypi.python.org/pypi/oct2py`, August 2015.

however, the focus is placed on the supervised learning content introduced in this chapter.

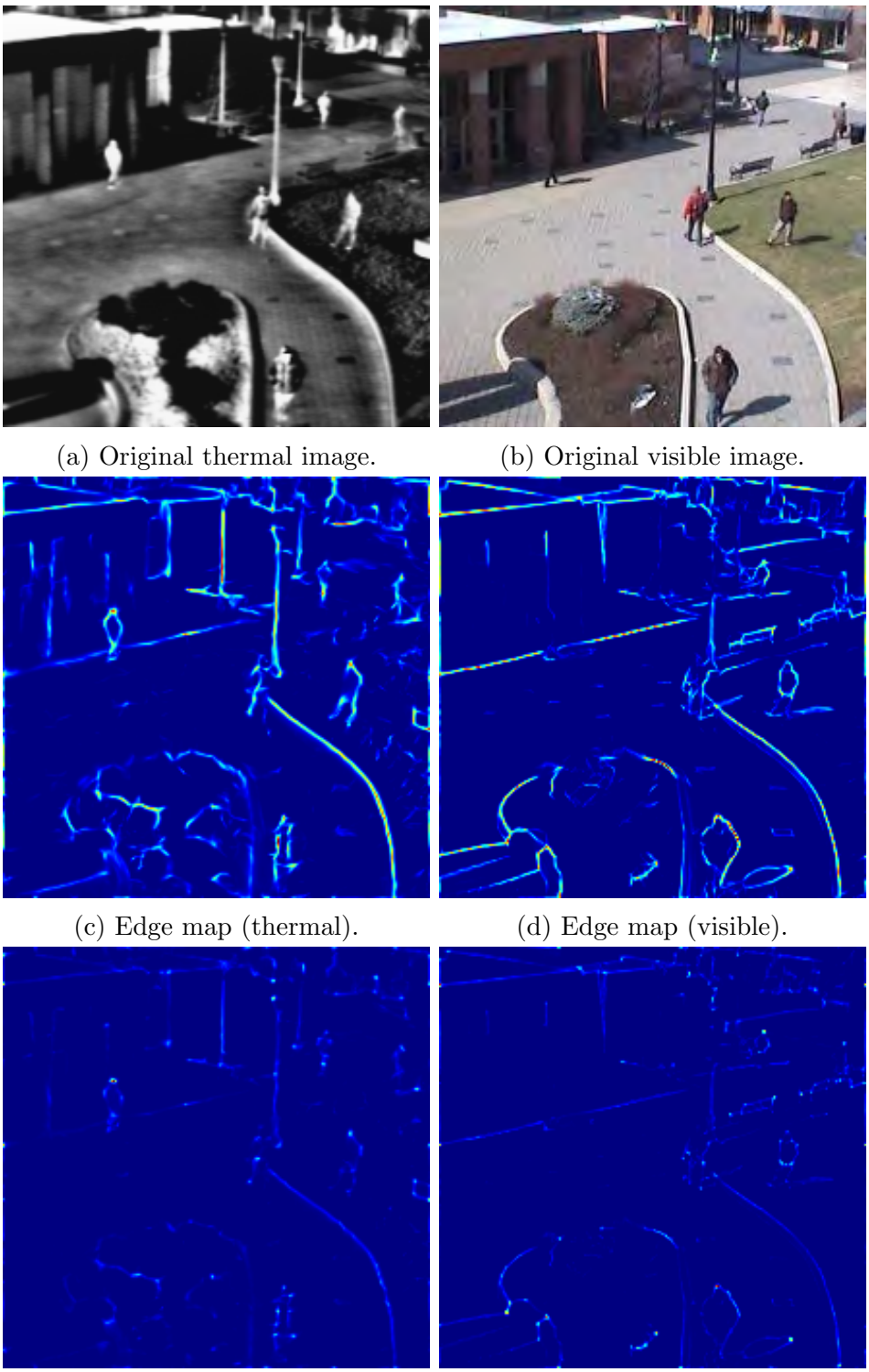
4.4.1 An invariant representation

A representation of the maximum and minimum moment values, which provide an edge and a corner map respectively, are shown in Fig. 4.16. The purpose of achieving an invariant representation is to enable methods that describe and match visible similarity to be used with multi-spectral information; therefore, the effectiveness of this representation can only be analysed in terms the performance it grants to these methods. The results presented here are the basis for the development work carried out in Chapters 5 and 6, and the evaluation of the invariant representation provided by phase congruency is deferred to its application in these chapters.

4.4.2 Predictable components of congruency

Structure of the investigation

The aim of this investigation is to identify a pattern vector that has a predictable relationship across the spectral modalities. Figure 4.17 provides a flow diagram of how the results are organised. The notation for each component is appended with an image identifier (i.e. 1 or 2) and directional filter orientation indexed by i (e.g. $\bar{\Phi}1_0$ represents the mean phase calculated with the directional filter-bank at orientation $i = 0$ from image 1). The components are initially partitioned into the oriented and reduced feature sets and analysed separately in blocks I and II; predictable components in each of these sets are identified and then combined in block III. Dashed lines indicate that the elements of the pattern vector were modified in some way (e.g. multiplied), and solid lines indicate that the pattern vectors were concatenated. Each transition is motivated by the performance, which is demonstrated with the ROC curve, offered by the new pattern vector.



(a) Original thermal image. (b) Original visible image.

(c) Edge map (thermal). (d) Edge map (visible).

(e) Corner map (thermal). (f) Corner map (visible).

Figure 4.16: The edge and corner maps of the reduced congruency feature set form the invariant representation of this project. (Modified images from the OSU database [44].)

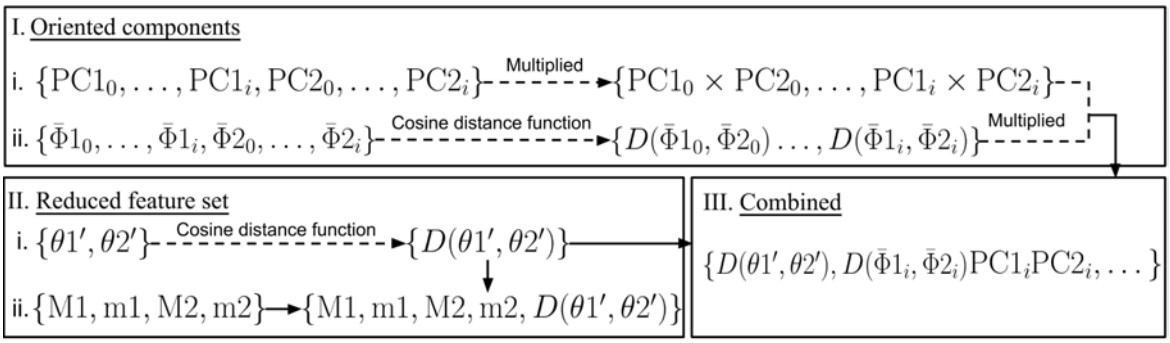
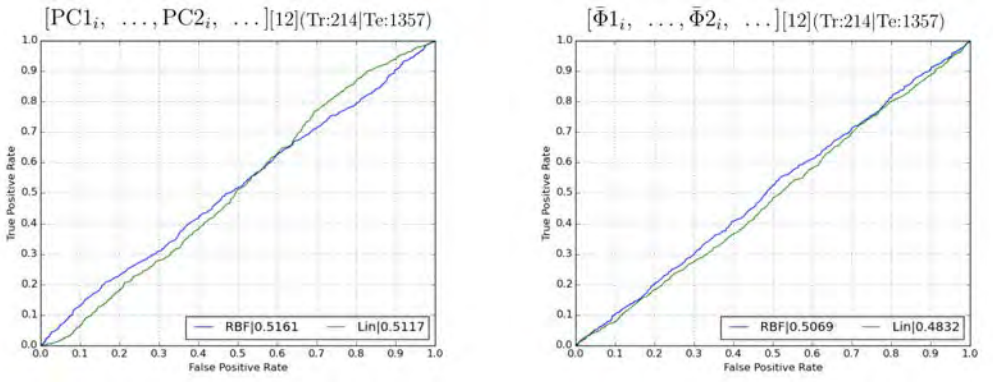


Figure 4.17: A flow diagram showing the structure of the investigation.

Oriented components

The oriented components of phase congruency, presented in Section 4.2.4, refer to the separate outputs of the six directional filter banks. The phase congruency and mean phase angles, introduced in Eq. (4.13) and Eq. (4.8) respectively, are used to form pattern vectors in the following experiments.

Figure 4.18 shows the ROC plots for the pattern vectors of the separate congruency and mean phase angles. The performance of both pattern vectors indicate no predictive value beyond random guessing (the ROC plots are along the diagonal).

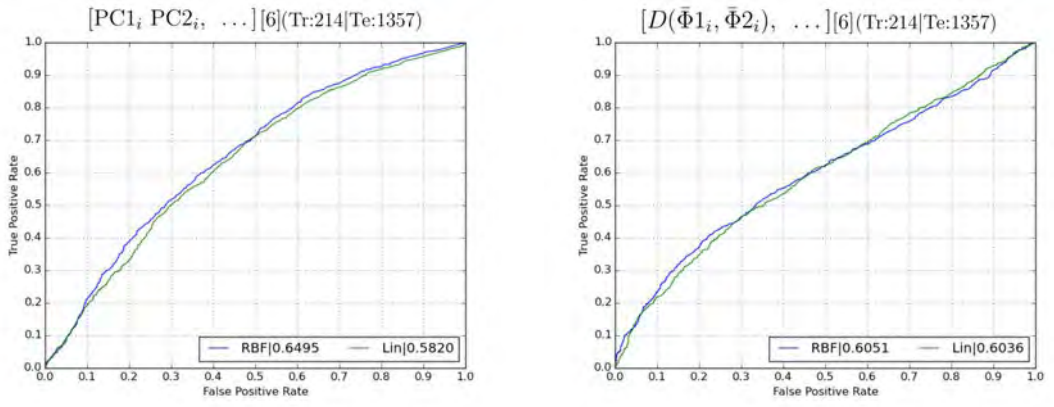


(a) Oriented congruency magnitudes PC_i. (b) Oriented mean phase angles $\bar{\Phi}_i$.

Figure 4.18: ROC plots of the twelve separate oriented components of the phase congruency magnitudes and mean phase angles (i.e. concatenation of outputs of six orientations from two images).

The result shown in Fig. 4.18 is used to demonstrate the performance offered by combining elements, shown as dashed lines in Fig. 4.17, reported in Fig. 4.21. The magnitude of oriented congruency PC_i provides an indication of structural significance (i.e. above the noise circle and consensus over multiple scales); for this reason, the

congruency at corresponding orientations is combined by multiplication to ensure that the resulting product is non-zero if the oriented congruency from both modalities is non-zero. The results are reported in Fig. 4.19a and show an appreciable improvement on the separate components of Fig. 4.18a. The angular mean phase components are combined using the cosine distance measure; however, only a small improvement over Fig. 4.18b can be seen in their combination in Fig. 4.19b.



(a) Corresponding congruency magnitudes PC_i are multiplied.

(b) The cosine distance is applied to corresponding mean phase angle values.

Figure 4.19: ROC plots demonstrating the performance of corresponding oriented components, extracted from each spectral modality, combined using multiplication (in the case of congruency) or cosine distance (in the case of phase).

The two results presented in Fig. 4.19 are combined by multiplying each respective oriented component; therefore, each component of the resulting vector is only large if the structures are similar (angular phase distance) and significant in both modalities. An improvement in prediction with the linear kernel using this combined pattern vector is shown in Fig. 4.20.

Reduced feature set

The second group of components investigated is the reduced feature set which was introduced in Section 4.2.5. The ROC curves of the moment magnitudes and feature orientation pattern vectors are shown in Fig. 4.21. Figure 4.21b shows that the feature orientation components are not linearly separable, but are separable with the RBF kernel. In order to make the components linearly separable, the cosine distance function is applied; the resulting ROC curve is shown in Fig 4.22.

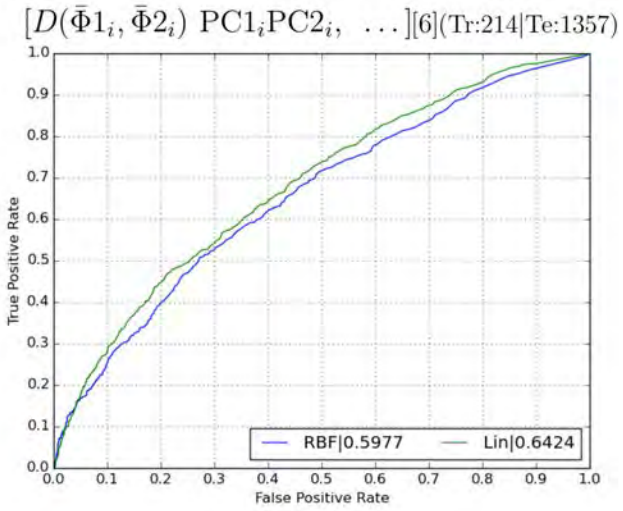
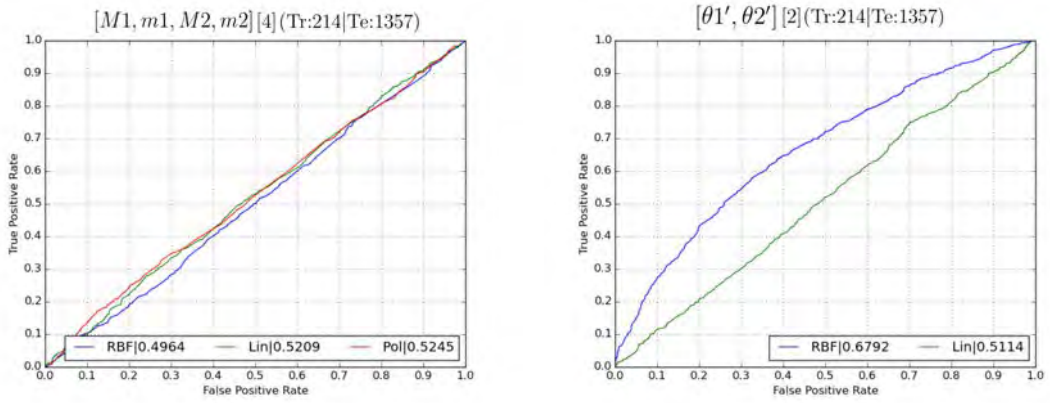


Figure 4.20: The oriented congruency and phase components are multiplied to form a single pattern vector. Each of the six (oriented) elements of the pattern vector represents the similarity of structure (phase angle) and weighted by the significance of the structure in both images (congruency magnitude).



(a) Maximum and minimum moment magnitudes.

(b) The angle of the maximum moment θ'.

Figure 4.21: Two ROC plots analysing the the separate elements of the reduced feature set. Plot (a) shows the performance of the maximum and minimum moment magnitudes, which indicate the presence of a feature. The pattern vector analysed in (b) consists of the edge orientations (angle of the principal axis) from each image.

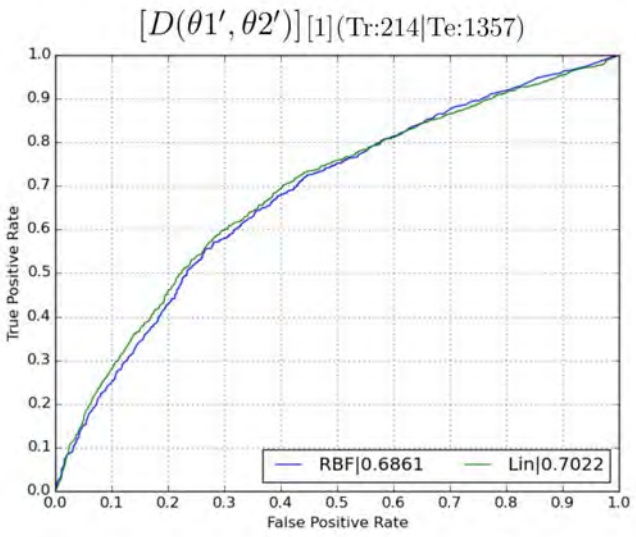


Figure 4.22: The angles of the maximum moment θ' are combined with the cosine distance measure to make the samples linearly separable. The linear kernel shows significant improvement (higher area under curve value) when the values are combined in this way.

The inclusion of the magnitude of the maximum and minimum moments showed no further improvement in performance over the of feature orientation distance only, and is not illustrated here.

Combination

The final step of this investigation, illustrated as block III in Fig. 4.17, was to combine the repeatable components identified in the results so far. The pattern vectors shown in Fig. 4.20 and Fig. 4.22 are concatenated and evaluated in Fig. 4.23. The components of the combined pattern vector describe a comparison of the features (i.e. feature orientation difference) and a structural comparison (i.e. distance and significance of oriented phase components).

The combined pattern vector in Fig. 4.23 demonstrates the best linear separability using seven components. The next section of this chapter provides a discussion on the practical implications of these results and how phase congruency can be used to identify and match corresponding observations from each spectral modality.

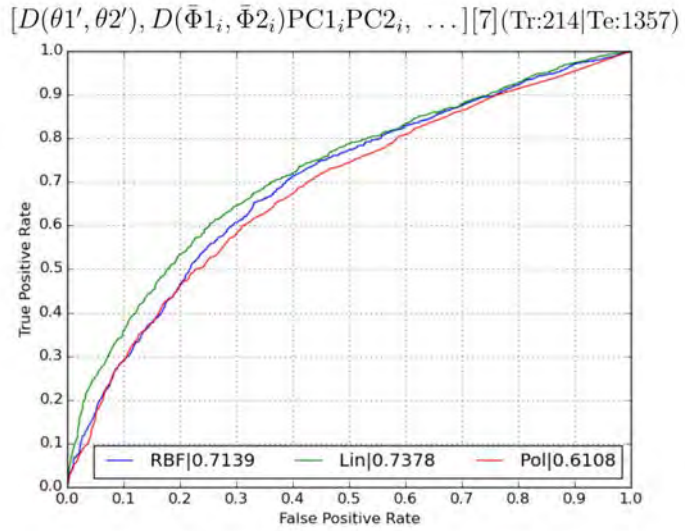


Figure 4.23: The edge orientation distance is concatenated with the six oriented phase distance and congruency measures to form a seven element vector. The linear kernel ROC curve shows the best performance over all the pattern vectors analysed in the investigation.

4.5 Discussion

The goal of this chapter was divided into two objectives. The first was to present phase congruency as a means of extracting an invariant representation of the common structural features from thermal and visible images. The second objective was to identify predictable features exposed by the phase congruency process between the two modalities.

Section 4.4.1 provided a cursory example of the invariant representation in the form of an edge and corner map in Fig. 4.16. An invariant representation is a visually similar depiction of the common information between the two modalities. The effectiveness of this invariant representation can only be quantitatively analysed when integrated into methods that use visual similarity to match regions. Chapters 5 and 6 both use the phase congruency edge map to adapt and enhance existing similarity metrics.

The second phase of this investigation, set out in Fig. 4.17, presented a series of experiments conducted to identify predictable pattern vectors exposed in the phase congruency process. Success was found by combining the scalar (significance) and angular components using multiplication and the cosine distance measure (seen in Eq. (4.20)) respectively. The final combined pattern vector, shown in Fig. 4.23, provided the best performance with a seven element vector; however, Fig. 4.22

demonstrated comparable performance with a single element pattern vector describing the angular distance between feature orientation values.

Description with a single element is very appealing as it can be easily integrated into a cost function. In order to further validate the effectiveness of feature orientation as means of describing structural features, a brief look at the distribution of values in the sample set is undertaken. The distribution of feature orientation of congruent edges in the sample set is shown in Fig. 4.24; the peak at $\pi/2$ is attributed to the prominent urban architectural features (i.e. vertical edges) in the images.

Figure 4.24: Normalised distribution of feature orientations of congruent edges in the sample set. The blue and red stepped distributions represent the feature orientation at edges extracted from the visible and thermal images respectively.

The cosine distance measure is then applied to corresponding samples to produce Fig. 4.25. The significant density of matches to the right of the histogram indicates that the measure is able to effectively identify similar edges. Feature orientation as an effective distance measure for rectified edge features is demonstrated in Chapter 6 of this work.

It is concluded that, while frequency domain analysis using directional log-Gabor filters does extract information with predictive value, the reduced feature set produced by Kovesi's phase congruency method provides invariant and predictable components that are better suited to integration with existing methods. The methods developed in later chapters of this work use the reduced feature set extensively to describe and compare regions of multi-spectral images.

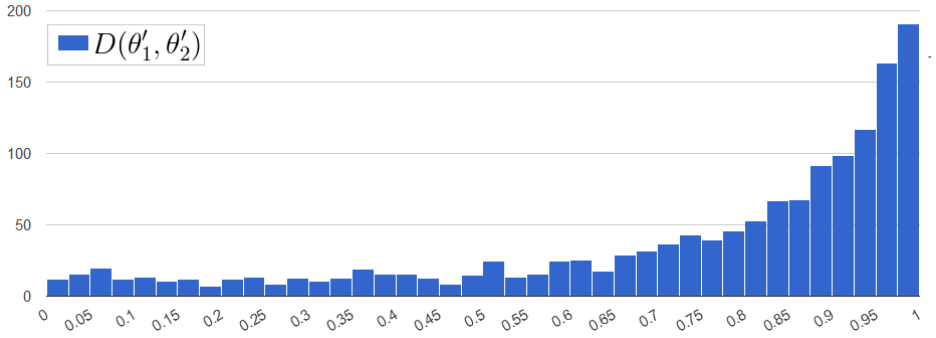


Figure 4.25: Histogram showing the distribution of values of the application of the cosine distance measure Eq.(4.20) to congruent edges in the sample set.

4.6 Summary

The investigation presented in this chapter motivates phase congruency as a means of exposing repeatable structure from thermal and visible images. Despite the non-linear relationship between pixel intensity values, presence of thermal phenomena and the low spatial resolution of thermal images, the method is able to extract repeatable and matchable structures in the two disparate modalities.

The next chapter presents an investigation into the feasibility of sparse correspondence methods for feature-based alignment of multi-spectral imaging modalities. The results of this chapter are integrated into a structural feature descriptor which uses the invariant structure exposed by phase congruency.

Chapter 5

Sparse correspondence methods

The majority of computational stereo methods require that the input images are rectified to make the challenges of stereo-vision more tractable. The rectification process, introduced in Section 3.2, brings the rows of each image into alignment by warping the image captured by each viewpoint. Although rectification is often accomplished through manual calibration, automatic alignment using sparse feature-based methods is commonly performed if the stereo-head is not reachable or if frequent re-alignment is required.

This chapter presents an investigation into the feasibility of conventional feature-based methods for automatic alignment of distinct multi-spectral viewpoints. Section 5.1 provides an overview of the terminology and role of conventional feature-based methods in viewpoint alignment, which is concluded with a brief discussion of the core assumptions of these methods. Conventional feature methods are reported to perform poorly (or fail entirely [91]) with multi-spectral images; a literature survey is performed in Section 5.2 to identify where and how these methods fail. The survey serves as a problem statement for the development of an adapted measure which incorporates the invariant elements of phase congruency presented in Chapter 4. Section 5.3 describes the approach to evaluating conventional and the adapted method. The results of this investigation are provided in Section 5.4, and are then discussed in Section 5.5.

5.1 Feature-based alignment

An image feature is a broad term referring to a point in an image that is deemed significant by some measure. For example, an edge is a structural feature usually indicated by a high gradient magnitude. This section introduces the traditional approaches to feature-based alignment through the detection and matching of corresponding observations from different viewpoints.

5.1.1 Terminology

The appearance of a scene can change drastically when viewed from a different position or under different lighting conditions; elements in a scene can undergo rotation, scaling and perspective distortion due to the change in viewpoint as well as illumination variations. These variations are collectively referred to as distortions in viewpoint or observation characteristics.

Interest points are sparsely-distributed points in an image that are stable (i.e. repeatably located) and matchable (i.e. containing descriptive information) under variable conditions. Methods called detectors localise these points in the image and often go further to capture the characteristic scale of the interest point [80]. A major criterion for detector performance is repeatability which refers to a detector's ability to consistently and accurately localise the same points in a scene regardless of variations in viewpoint characteristics.

The local region around an interest point is described, or encoded, by a numerical feature vector which can be used to gauge the similarity between interest points. Descriptor methods construct this feature vector such that it can be matched between corresponding viewpoints even if the local region has become distorted between observations. Feature vectors are compared using a distance measure that provides a numeric value representing the confidence that the two vectors correspond to the same observed point.

These stages are referred to as detection, description and matching and are applied sequentially to identify correspondences. Although these stages consist of distinct methods, the majority of detectors are optimised to identify structures tailored to the descriptor and are therefore not practically interchangeable.

5.1.2 Conventional approaches

Detection and matching of sparse image features is vital in many applications and is the conventional approach to automatic alignment of images taken from different viewpoints. Numerous methods of feature detection and description have been presented in literature; this section presents a representative sample of common approaches to this task.

Many of the earliest detection methods identified corners as stable interest points. The robust Harris corner detector [81] was first presented in 1988 as a refinement of existing approaches to corner detection [82]; it identified corners as points in an image where an autocorrelation function was maximised. The Shi-Tomasi corner detector (or the Good Features (GF) corner detector from the title of the paper "Good Features to Track") provided further refinement to this method by reducing the frequency of edge responses, and is an important part of the investigation later in this chapter. Also included is the modern Features from Accelerated Segment Test (FAST) [83] corner detector which is commonly used with the Oriented Fast and Rotated BRIEF (ORB) [84] and Binary Features from Robust Orientation Segment Tests (BFROST) [85] binary descriptors.

Binary descriptors have recently received a large amount of interest due to the computational and memory limitations of mobile devices [84, 86]. Binary descriptors are compact binary strings formed by pair-wise intensity comparisons of pixels radially distributed about an interest point; each bit in the descriptor represents the logical outcome of an inequality between the intensity values of each pair. The Binary Robust Invariant Scalable Keypoint (BRISK) [87] descriptor operates on corners detected with the Adaptive and Generic Accelerated Segment Test (AGAST) [88] detector to provide rotation and scale invariance. Binary strings are compared with the Hamming distance: each element is compared with a logical XOR operation, and the sum of the resulting bits represents the binary difference [86].

Possibly the most common approach found in literature and in practice is Lowe's Scale invariant feature transform (SIFT) [89] method of detection, description and matching [90]. SIFT is prominent in literature as it has become a robust benchmark against which the performance of novel methods can be measured [86]. The detection stage of SIFT localises blob-like interest points as maxima in scale-space in a Difference of Gaussian (DoG) image pyramid. The descriptor itself is a 16×16 pixel window in

which each pixel contributes a vote, weighted by gradient magnitude, to a histogram of discrete orientations used to describe the region. The notable Speeded Up Robust Features (SURF) [90] method is based on the SIFT descriptor and provides an efficient alternative at the cost of matching performance. Both SURF and SIFT feature vectors are compared with the L2-Norm measure, a floating-point arithmetic operation.

The use of gradient magnitude as a significance measure is seen in the majority of detector and descriptor algorithms. While its use in SIFT to weight votes is explicit, the pair-wise intensity comparisons of the binary descriptors are essentially encoding the sign (i.e. increasing or decreasing intensity) of the gradient. The assumption that gradient direction remains constant is fundamental to describing elements in visible spectrum images; however, the non-linear relationship between visible and infrared modalities breaks this assumption. This chapter presents an investigation into the performance of conventional feature-based methods in a multi-spectral context and proposes an alternative means of describing image elements that is invariant to the variations between the spectra.

5.2 Multi-spectral image features

Traditional approaches to sparse feature correspondence operate under assumptions that hold within the visible spectrum. This section presents a brief background, the purpose of which is to determine where traditional methods fail, identify the practical limits of their application and establish a context to the investigation carried out in this chapter. A survey of the performance of conventional feature-based methods reported in literature is presented in Section 5.2.1. Observations from this survey are used in Section 5.2.2 to formulate the objectives of the investigation, and in the development of a multi-spectral descriptor in Section 5.2.3.

5.2.1 Problem statement

There is a clear consenus that traditional feature correspondence methods perform poorly or fail entirely to detect and match thermal and visible features [38, 43, 91], although there has been limited investigation focused on analysing this behaviour beyond the visible spectrum [18, 23].

In order to identify where traditional methods fail, start with the first assumption that these methods are only suitable for use with visible spectrum images. To dispute this, Ricaurte et al. in [18] provides an extensive investigation of the intra-band behaviour of SIFT, SURF, BRISK, Binary Robust Independent Elementary Feature (BRIEF) [92], ORB [84] and Fast Retina Keypoint (FREAK) [93] algorithms on thermal image pairs compared to their visible spectrum performance of the same scene. While noticeably weaker than operation in the visible spectrum, the traditional methods (SIFT, in particular) performed well enough in the thermal spectrum to warrant practical use. This intra-spectrum performance is used in applications presented in the literature [24].

The second logical assumption is that traditional methods will only perform well within the same spectral range. This is partially true; minor adaptations on the traditional SIFT descriptor have been required to enable VS-NIR feature matching [22]. However, as the difference of wavelength between the spectral modalities increases, the images become increasingly dissimilar and traditional feature-based correspondence methods fail entirely [23]. This failure is attributed to the low spatial resolution, noise and the lack of common descriptive elements (i.e. colour or texture) in thermal images, as well as the increasingly disparate information captured by the distinct sensors.

The result of this dissimilarity is that feature detectors do not locate the same observations and feature descriptors that rely on gradient direction fail due to the non-linear relationship between the intensity values of each spectral modality [91]. Contrast reversals are common in the distinct observations captured by these sensors; for example, the edges of a dark, warm object on a light, cool background would appear highly dissimilar to SIFT due to the gradient directions being 180 degrees out of phase (i.e. contrast is reversed). Numerous adaptations of the SIFT algorithm have been made to compensate for and correct contrast reversal (e.g. Gradient Direction Invariant SIFT (GDI-SIFT) [22], Oriented SIFT (OR-SIFT) [94], Uniform Robust SIFT (UR-SIFT) [95]); however, these methods are designed for matching visible with near infrared images and not suited to matching across the large spectral gap between visible and thermal information [23, 43].

5.2.2 Hypothesis and objectives

The goal of the remaining sections of this chapter is to present an approach to multi-spectral feature detection and description that is better suited than traditional

methods. This goal is based on two hypotheses: first, that traditional approaches are fundamentally unsuited to matching multi-spectral features, and second, that the detection and description of structural features (such as edges and corners) provide a more repeatable and matchable similarity measure than gradient-based methods.

The specific objectives are therefore to develop and present a set of methods that detect and describe structural features such that they can be matched and to demonstrate the effectiveness of the approach by comparing matching performance to traditional methods.

5.2.3 Development of theory

Object and material boundaries manifest as edges and ridges and are often described in terms of their orientation in the image. These structural features are commonly detected at pixels with a high gradient magnitude and described in terms of the direction of the gradient, perpendicular to the edge. The method proposed in this section has close parallels to the work of Park et al. [96], and was closely influenced by the work of Aguilera et al. [35, 38] and, in particular, Mouats and Aouf [43].

Conventional methods rely on gradient magnitude to indicate significance in detection and gradient orientation as a means of description; however, the disparate information captured by thermal and visible sensors means that these assumptions, based on intensity information, do not hold in most cases. Methods that utilise frequency-domain analysis for detection and description are proposed in this work as a means for repeatable detection and inter-spectrum feature description. Phase congruency, introduced in Chapter 4, is a recurring focus of this work. It is used here as robust corner and edge detector, invariant to image contrast, to replace the gradient magnitude approach of traditional methods.

A variation of the Edge Orientation Histogram (EOH) descriptor is developed in this dissertation. The EOH descriptor, as presented in [96], is predominantly used in the MPEG-7 standard [97]. It is noted for its efficiency and efficacy in image content search and retrieval, and it is often applied as a global or semi-global descriptor in order to retrieve similar images from a database [96–98]. This class of edge orientation methods influenced the Histogram of Oriented Gradients (HOG) [99] descriptor which has been particularly effective in hand gesture recognition, a task requiring invariance

to photometric qualities such as illumination, skin colour and background colour when comparing an observed gesture to a database.

The EOH feature vector is constructed with an edge orientation voting process in which the orientation of each edge pixel is used to construct a histogram of orientation votes that describe the local region. The square $N \times N$ pixel descriptor footprint is divided into 16 sub-regions; each sub-region holds a voting process to construct a histogram of the four directional 0, 45, 90, 135 degree bins and the no orientation bin. The histograms from each sub-region are separately normalised and concatenated to form the 80 element EOH feature vector (i.e. 16 sub-regions each contributing a 5 bin orientation histogram). An illustration of this process is shown in Fig. 5.2.

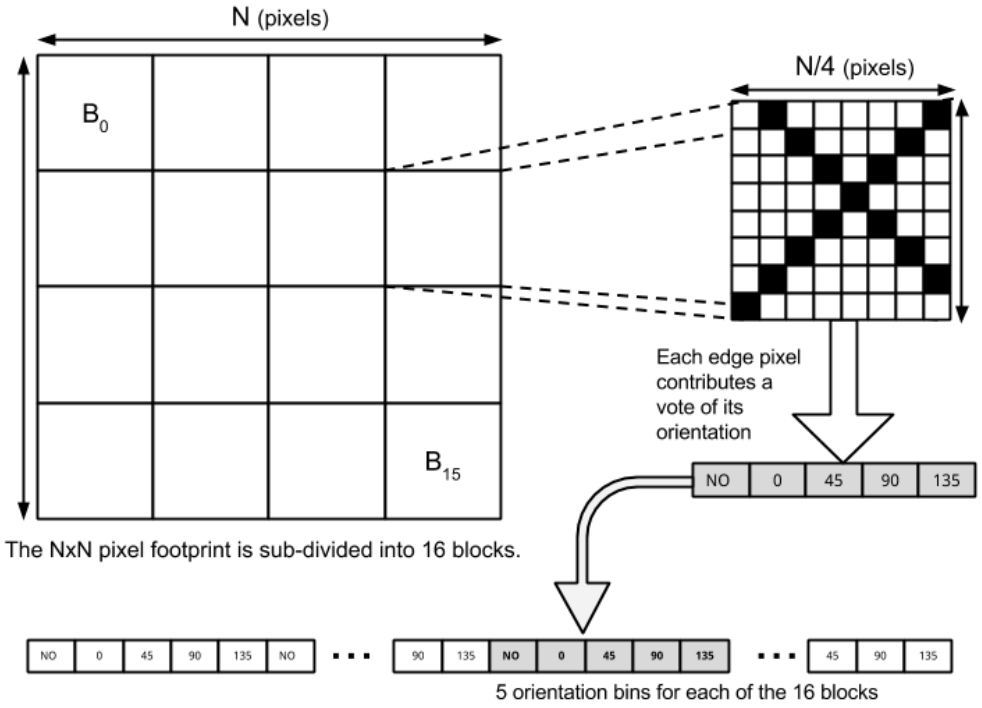


Figure 5.1: The edge orientation histogram (EOH) is an $N \times N$ pixel square patch divided into 4×4 blocks, each characterised by five bins containing the votes for the gradient direction of its pixels [35].

Edge orientation voting is a task common to many descriptor algorithms; the SIFT descriptor, for example, uses 36 orientation bins. The vote of each edge pixel of the EOH descriptor is based on which of the five 3×3 spatial filters is maximised. Figure 5.2 illustrates two commonly used spatial filter configurations. The numerical values in each cell represent coefficients assigned to pixel values in the convolution operation with the underlying edge map.

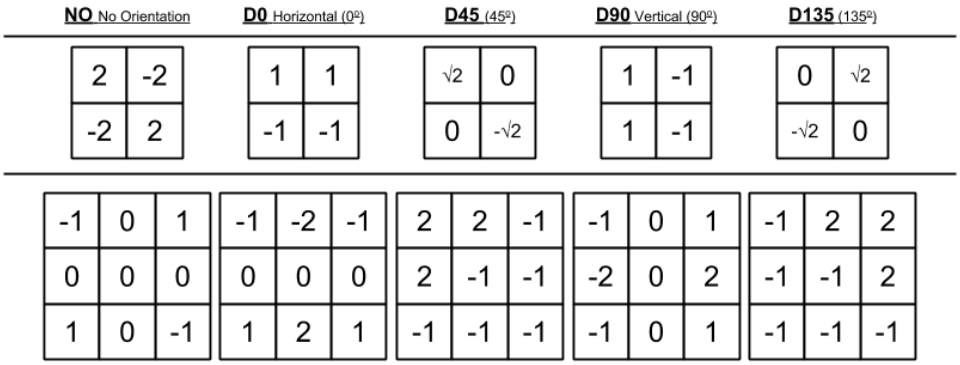


Figure 5.2: The directional filters shown above are applied to each pixel in the EOH. Edges are classified into four discretised directions, with an additional "no orientation" filter. The first row shows the 2×2 filters used by Park et al. [96], the second row shows those used by Aguilera et al. [35].

5.3 Approach to evaluation

This brief section presents the evaluation of conventional feature-based methods and the EOH descriptor method developed in the previous section. The approach to evaluating the different stages of feature-based methods is described in Section 5.3.1, and the tools (i.e. algorithms, software libraries and inputs) used are recorded in Section 5.3.2.

5.3.1 Quantifying performance

In order to achieve the objectives set out for this investigation, feature correspondence methods are split into detection and matching stages and analysed separately where possible.

Interest point detectors are analysed based on detection *repeatability*: the ability of a detector to identify and localise the same observations in the two viewpoints. Detecting the same points in both viewpoints is a crucial first step as it generates the pool of matchable points. Detector repeatability is evaluated to determine how many of the interest points detected in one modality are also found in the other; it is represented as a fraction of the simultaneously observed interest points over the total number of feature points detected. In order to avoid the use of a heuristic threshold on each detector's interest measure, a constant number of the strongest interest points across

all detectors is taken. While this constant is also heuristically chosen, it standardises the way in which the detectors are evaluated.

Interest point matching demonstrates a descriptor's effectiveness in encoding the information at interest points such that it can be correctly matched. Evaluation of descriptors is a little more complicated in that a number of steps must be taken to standardise the matching process. Once two sets of feature vectors have been extracted from the images, a simple brute force matching algorithm is used: each feature vector of one set is compared to every feature vector of the other set and associated (matched) with the vector that optimises the similarity metric. In this work, the declared matches of the descriptors are limited to the matches that have a distance less than twice the best (smallest distance) match. Descriptor performance is evaluated based on the number of true positives (correctly identified as a match) and false positives (incorrectly identified as a match) within the declared set of matches for each descriptor.

5.3.2 Tools and parameters

In order to achieve the objectives set out for this investigation, a representative sample of detectors and descriptors was chosen. The stand-alone detectors evaluated are the Shi-Tomasi and phase congruency corner detectors, as well as the detectors of the SIFT, SURF and FAST methods. The conventional SIFT, SURF and FAST algorithms were then analysed (with their respective detection methods) alongside the EOH descriptor (with the Shi-Tomasi corner detector) using a brute force matching algorithm.

The algorithms (with the exception of the EOH descriptor) were implemented using the OpenCV library and the Python programming language. The parameters suggested by the respective authors of each method were used. The EOH descriptor uses the phase congruency tools, described in Section 4.3.4, and the numpy library for implementing efficient linear filtering and edge orientation voting. Automated analysis was done using two image pairs from the OSU (OCTBVS) Color-Thermal Dataset with a ± 2 pixel ground truth. Although the corresponding points were removed from the two sets once matched, clustering of interest points resulted in the detection results not always being symmetric across the spectral modalities. A maximum of 200 interest points was taken from each image in the analysis of detector repeatability, although this was increased to 300 when descriptor performance was analysed in an effort to increase

the number of potential matchable points. This maximum is heuristically chosen, but it helps to standardise comparison of the detection and matching algorithms.

The proposed EOH descriptor was then investigated under a maximum feature displacement constraint of ±40 pixels coupled with outlier rejection with Random Sample And Consensus (RANSAC) [80] (OpenCV implementation). These constraints are commonly used in practice [43, 100] and are used here to better characterise the practical performance of the proposed method. In order to demonstrate its performance on a different dataset, a sample from the CVC Multimodal Stereo Dataset (2) was manually evaluated and included.

5.4 Results

Of the detectors analysed, the Shi-Tomasi GF corner detector performed the best on both sample pairs, providing 35% repeatability from visible to infrared. Despite showing similar performance in one case, the phase congruency detector suffered due to a large percentage of detected interest points clustering along different edges in the two spectral modalities. Due to the 3 pixel non-maximal suppression of these methods, the results were symmetric. The FAST detector demonstrated repeatability of thermal interest points in the visible image (i.e. a high percentage of interest points detected in the thermal image were found in the visible image), but showed very poor repeatability of visual interest points in the thermal image. The SIFT and SURF interest point detection methods performed very poorly; less than 10% of interest points were common to both viewpoints.

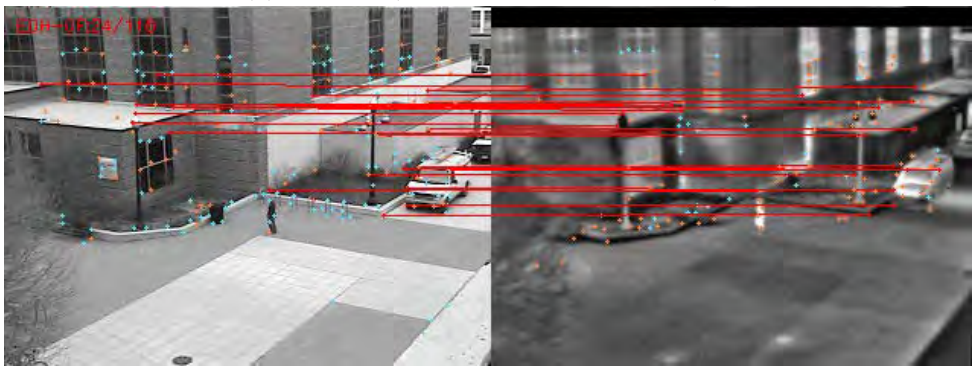
The conventional SIFT, SURF and BRISK detection and description algorithms were matched using the methods described in the previous section. SIFT matched 4.3% (4 correct, 94 declared) as did SURF (3 correct, 69 declared). BRISK performed the worst at around 1.2% (3 correct, 249 declared). Under the same conditions, the EOH-GF algorithm matched 18.1% (39 correct, 215 declared). Figure 5.3 illustrates one such case to emphasise the performance of the EOH-GF algorithm.

The EOH-GF method was analysed with a maximum displacement constraint (±40 pixels) and RANSAC outlier rejection. The combined results showed that 25% of the interest points (115 of 460 detected) were common to both modalities. Brute force matching resulted in 34% (84 of 247 declared) of features being correctly

(a) SIFT (1 correct of 46 declared).

(b) SURF (1 correct of 60 declared).

(c) BRISK (1 correct of 113 declared).



(d) EOH-GF (24 correct of 116 declared).

Figure 5.3: The inter-spectrum (visible to thermal) matching performance of the SIFT, SURF, BRISK and EOH-GF methods are illustrated in four image pairs. Correct matches are joined by red lines, while false positives and detected (but unmatched) interest points are shown as orange and blue crosses respectively. (Images from the OSU database [44].)

matched, which increased to 58% (51 of 88 declared) after RANSAC was used to reject outliers. Figure 5.4 demonstrates the performance of the EOH-GF algorithm (under the constraints discussed) on a higher resolution thermal/visible image pair. The image pair is rectified although no ground truth exists; therefore, correspondences were manually analysed and circled in green (correct) or red (incorrect).

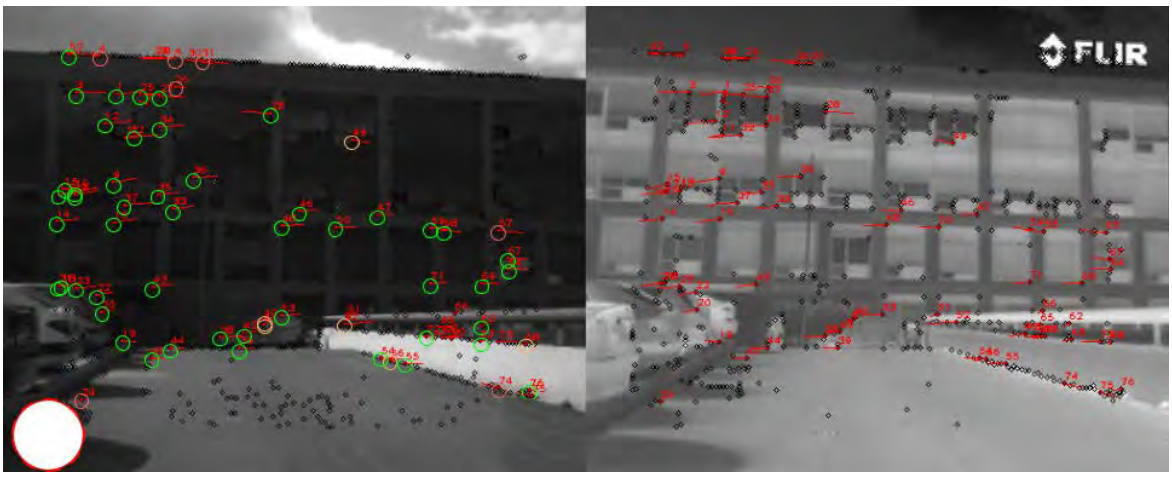


Figure 5.4: Of the RANSAC-declared EOH-GF 78 matches, 53 are correct (within 3 pixels) and 10 are incorrect. There are 4 matches of adjacent features and the remaining 7 are difficult to discern. The white circle on the bottom left hand side of the image indicates the extent of the disparity constraint for each match. (Image from the CVC dataset [34, 35].)

5.5 Discussion

The results show that conventional methods for feature detection and description are fundamentally unsuited to matching between multi-spectral images. The derivative-based SIFT and SURF detectors rely on stable regions of constant intensity (often called blobs) to localise features in scale-space; however, these structures seldom simultaneously occur in both visible and thermal observations of a scene. The descriptors are rendered useless as they fail to accommodate for the non-linear relationship between the modalities. This poor performance in detection and description is exacerbated by the low spatial resolution and noise of the thermal sensor.

The results of the conventional SIFT, SURF and FAST methods are consistent with the findings of [23, 35, 91]; however, the performance of the Shi-Tomasi corner detector seems at odds with the findings of Han et al. [91]. A brief experiment showed that

the Shi-Tomasi corner detector produced different feature points with the OpenCV implementation used in this work when compared to the MATLAB implementation used in [91], which may explain the difference in results.

The repeatability demonstrated by the Shi-Tomasi GF corner detector in this work is attributed to three factors. Firstly, non-maximal suppression ensures that a single point is selected for each corner and reduces ambiguous matches by increasing the spatial distribution of points. Secondly, selecting a constant number of features and not exclusively corners with high responses reduces the reliance of the detector on gradient magnitude and heuristic thresholding. Finally, the Shi-Tomasi GF corner detector is designed to reject edge responses, resulting in more corners being found despite the presence of strong edges in the images. The frequency-based detection of the phase congruency corner detector was particularly hampered by clustering along edge features.

Despite its simple implementation, the EOH descriptor significantly outperformed the conventional methods of feature point description. The results demonstrate that detection and description based on structural features (i.e. corners and edges) provides repeatable performance and encodes stable elements of multi-spectral images. As the EOH descriptor essentially compares visual similarity, this result also shows the efficacy of phase congruency in extracting a stable invariant representation of multi-spectral image data.

The 58% correct matching rate is very close to the 56% achieved by the similar implementation by Mouats and Aouf [43] which uses the same dataset, but performs analysis over a greater number of frames to track pedestrians moving across the static scenes. Only two images were used in this study (one pair from each sequence) because analysis of the predominantly static features was deemed to not add significant value to achieving the aims of this chapter. There are two significant differences in implementation: the different corner detector (Shi-Tomasi instead of phase congruency) and the smaller descriptor size (40×40 instead of 100×100). The smaller detector footprint is more sensitive to local changes and is therefore more distinctive. Furthermore, using a 100×100 pixel on thermal images with 320×240 pixels defeats the purpose of using local features, which are intended to provide a means of efficiently matching sparse correspondences of significant features.

Although the proposed EOH-GF descriptor demonstrates significant improvement on conventional methods, the conclusions drawn from these results lack weight due to the

limited dataset available. To alleviate this, the descriptor was applied and manually evaluated on samples of a different dataset (an example is seen in Fig. 5.4); however, further automated evaluation should be undertaken when datasets become available.

Despite the improvement over conventional methods in handling inter-spectrum variations, this descriptor is still limited in its applicability to real-world image alignment. The performance of the descriptor drops off significantly — to the point of being unusable — when faced with rotation, scaling and perspective distortion present in unaligned images. While its sensitivity to these factors can be decreased by a larger spatial foot-print, its lack of scale invariance makes it unable to deal with the significant difference between the spatial resolution of visible and thermal cameras. Further development of both the detection and description stages of the EOH method is required to achieve level of scale and rotation invariance needed for matching between unaligned images.

5.6 Summary

This chapter addresses the task of aligning two images from thermal and visible spectral modalities. The most common approach to multi-view image alignment is to automatically identify correspondences in a process consisting of detection, description and matching stages. The significant improvement over traditional correspondence methods was demonstrated in the development and evaluation of the EOH-GF descriptor; however, it was concluded that further of development and evaluation would be needed to establish the method as a practical structural descriptor for unconstrained multi-spectral image alignment.

The next chapter presents the development of a cost function for multi-spectral stereo correspondence matching. The goal of the chapter to construct a disparity map in order to address the second objective of this work, i.e., to overlay multi-spectral image information. It will be assumed that the two viewpoints have been rectified using one of the processes described in Section 3.2.

Chapter 6

An invariant similarity measure

Comparing points or regions of thermal and visible spectrum images is challenging due to the non-linear and uncorrelated relationship between the brightness and contrast of the modalities. Stereo correspondence methods, which were introduced in Section 3.2, estimate the depth of objects in the scene by calculating the horizontal displacement of points in the scene projected onto each image plane. However, this process is complicated by the large textureless regions and low spatial resolution, typical of thermal images, coupled with non-simultaneous phenomena (observations apparent in one spectral modality only).

This chapter introduces a block-matching method for generating sparse correspondence across aligned thermal and visible images; the goal is to develop a cost function for computational stereo methods that enable us to overlay regions for multi-spectral information fusion. Section 6.1 provides a background to mutual information and its applications in image registration, although the focus is placed on the use of mutual information as a local region descriptor. A survey of adapted local measures of mutual information is provided in Section 6.2. The investigation carried out in this chapter aims to assess suitability of mutual information for computational stereo and multi-spectral region matching, and to determine the effectiveness of incorporating components of phase congruency in the measure. Specific objectives and hypotheses of the investigation are clarified in Section 6.3. The approach to achieving these objectives is described in Section 6.4. The results of the investigation are then presented in Section 6.5, followed by a discussion of the findings and a brief look at the practical implications of the work to conclude the chapter.

6.1 Mutual information

Mutual information is a non-linear similarity measure that has been shown to be highly effective as a multi-spectral distance measure, outperforming traditional stereo correspondence measures [101]. This proven track record has been a motivation for its inclusion in many works [47, 102, 103]. It is of particular interest in this dissertation as it has been shown to successfully match regions of different spectral modalities; the measure relies on few assumptions about the underlying data and is able to adapt to the highly variable content of multi-spectral images [104].

The following section presents a brief background to mutual information and motivates its use as a method for matching regions of thermal and visible images.

6.1.1 Background

Woods [105, 106] is credited with the initial development of mutual information for medical image registration [8]. Multi-spectral medical imaging technologies capture complementary information about a patient that can be fused to aid human interpretation towards a diagnosis [7]. Mutual information stems from studies of entropy and information in communication theory [8, 104], although the bulk of literature on mutual information in computer vision is in the context of medical imaging [7, 8].

Mutual information is a tool used to describe the statistical dependence between two signals, allowing us to quantify how much information one signal communicates about the other [107, 108]. Consider two signals, A and B, each quantised to contain a finite and equal number of symbols. Both signals are simultaneously sampled, i.e., for each sample $a_i \in A$ there exists a corresponding value $b_i \in B$. Mutual information, $I(A, B)$, uses this extra information to reduce uncertainty in future samples by inferring a statistical relationship between the paired samples of the two signals; Figure 6.1 illustrates this concept using a Venn diagram.

Entropy for signals A and B is calculated independently and most commonly using

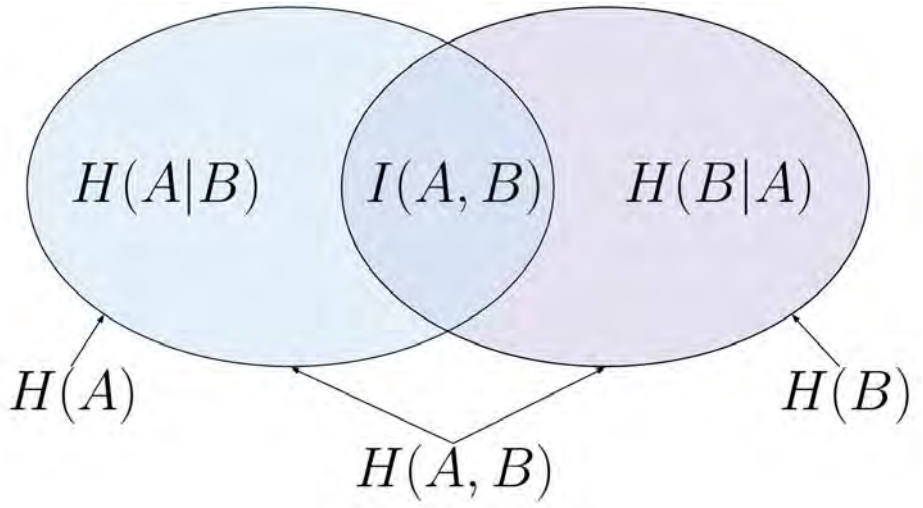


Figure 6.1: Venn diagram showing mutual information, $I(A, B)$, as the intersection of the information of A and B [109].

the Shannon entropy measure [7, 8]:

$$H(A) = -\sum_i P(a_i) \log_2 P(a_i) \tag{6.1}$$

where $P(a_i)$ is the probability of the symbol a_i in the signal A. Pluim et al. [110] describe entropy as a measure of the dispersion of a probability distribution; entropy is maximised when all symbols have an equal probability of occurring (a maximally dispersed uniform probability distribution). Figure 6.1 illustrates the reduced uncertainty in the conditional entropies $H(B|A)$ and $H(A|B)$ due to the presence of the other signal being simultaneously sampled. Interpreting the Venn diagram, one can say that, given another sample b_i, we can predict the corresponding a_i with expected uncertainty reduced by $I(A, B)$. In the context of mutual information, the entropy of the two sources $H(A)$ and $H(B)$ are called the marginal entropies of the calculation.

The following equivalent formulations of mutual information can be drawn from the Venn diagram in Fig. 6.1 [8]:

$$I(A, B) = H(B) - H(B|A) \tag{6.2a}$$

$$I(A, B) = H(A) + H(B) - H(A, B). \tag{6.2b}$$

The first definition reiterates the description of mutual information provided so far: mutual information is the amount the uncertainty about image B decreases when the relationship $a_i \rightarrow b_i$ is known. The second definition incorporates the joint entropy

function $H(A, B)$ which is a two-dimensional joint probability function representing the mapping of intensity values between the two signals [108]. Maximisation of mutual information is traditionally done by finding the image transform that minimises this joint entropy function.

Figure 6.2 provides an example of the joint histogram of a correctly registered '**T**' shape. The example demonstrates how the pixel intensity mappings create peaks in the joint entropy histogram (on the right) when the images are correctly aligned. Joint histograms with many intensity levels (e.g. 2^8 for grey-scale images or 2^{14} for thermal images) are visualised as intensity images in which each pixel represents the probability of that particular $a_i \rightarrow b_i$ mapping; the joint probability histogram of correctly aligned multi-spectral images contains clusters of common mappings.

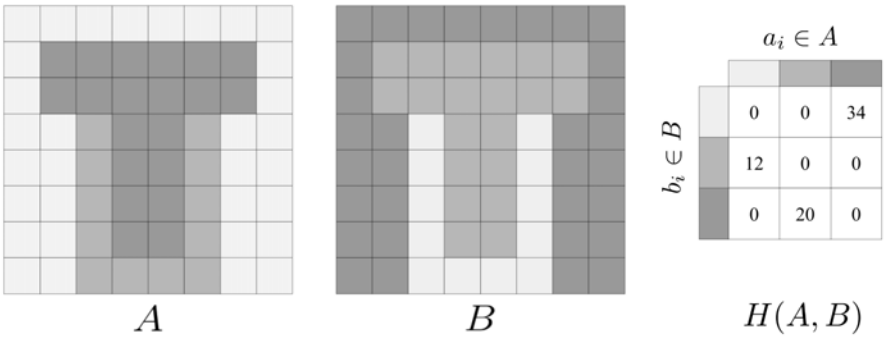


Figure 6.2: Illustration showing two '**T**' shapes A and B with different intensity levels. The right most image shows joint entropy histogram, illustrating the frequency of intensity value mappings between A and B. The overlayed shapes create peaks in the joint entropy histogram, minimising the joint entropy $H(A, B)$ (and thereby maximising the mutual information) between the two sources. This image illustrates how mutual information can be used to register regions which are not linearly related.

Alignment of calibrated medical images involves estimating a rigid transform that globally aligns the two input images; this is often done by a human operator who, by manipulating the images to minimise the perceived (visual) dispersion of the two-dimensional joint probability histogram, minimises the joint entropy value. The presence of known prominent structures in the images (e.g. bone, cartilage or tissue) creates common mappings which appear as bright clusters in the joint entropy histogram; these clusters are distinctive enough to allow a human operator to accurately align the images.

6.1.2 Local region similarity with mutual information

In this work, mutual information is used as cost function in a block-matching method for computational stereo, which was introduced in Section 3.2. The horizontal displacement of a point is estimated by comparing the local region of the point to regions along the corresponding row in the other image and selecting the point that maximises the cost function.

The observed disparity of a point is the horizontal pixel displacement between the projections in the image planes of the distinct viewpoints. Disparity is estimated by selecting a point/region, A, in one image and iteratively comparing it to regions, B, along a search domain of the possible disparity values. Therefore, each comparison will involve a different marginal entropy $H(B)$ as the comparison window moves along the corresponding row. With reference to Eq. (6.2b), it can be seen that if the marginal entropy of B increases faster than the joint entropy function is minimised, the mutual information metric will be maximised despite a decreasing overlap between the two frames. To compensate for the change in marginal entropy, a normalised mutual information measure [8, 110]

$$Y(A, B) = \frac{H(A) + H(B)}{H(B, A)} \tag{6.3}$$

is used; this adapted measure is investigated in this report.

Local regions of natural scenes do not contain the significant structures that are typically available in medical imaging. The next section describes how the performance of mutual information can be improved by incorporating spatial information and explains how the two measures are combined.

6.2 Spatial information

Mutual information is a statistical measure that does not take spatial information into account (i.e. the value of each pixel in the context of its neighbours.) Mutual information often produces many local maxima when the region is small or untextured [8]. To reduce the number of ambiguous matches within the search domain, mutual information is often combined with other distance measures such as

gradient [2, 37, 110] and local phase [46, 74].

Pluim et al. [110] presents the combination of mutual and gradient information in the context of medical imaging. Gradient information is defined as

$$G(A,B) = \sum_{(x,x')\in(AB)} w(\alpha_{x,x'}(\sigma))\ \min(|\Delta x(\sigma)|, |\Delta x'(\sigma)|). \tag{6.4}$$

The term $w(\alpha_{x,x'}(\sigma))$ is the familiar cosine distance function, described in Eq. (4.20), applied to the three normalised gradient vectors obtained through convolution with Gaussian partial derivatives at scale σ. It is noted that large gradients appear at tissue boundaries, but not all boundaries are present in both modalities. Therefore, a weighting function, $\min(|\Delta x(\sigma)|, |\Delta x'(\sigma)|)$, is applied to the cosine distance to ensure the presence of edges in both modalities. Barrera et al. [2, 37] use the same gradient calculation, but increase robustness by propagating both gradient and mutual information through a coarse-to-fine Gaussian pyramid.

This work incorporates the reduced feature set of the phase congruency process discussed in Chapter 4. Local regions extracted from the edge (the maximum moment M) maps of each modality are supplied as the inputs to the mutual information equation. Spatial information is incorporated with the edge orientation θ' value, and corresponding pixels are compared with the cosine distance measure; however, the angular distance between the corresponding points in each region is only considered significant if the value of the maximum moment of both points is non-zero. Although this approach is similar to the gradient method reported in the previous paragraph, the angular distance is *not* weighted by the edge magnitudes. This decision is based on the results of the investigations of Chapter 4 where there was found to be no relationship between corresponding edge magnitudes extracted from each modality.

The mutual information and feature orientation values are combined, equally weighted, by multiplication [2, 37, 110]; consequently, maxima only occur if both measures agree. Figure 6.3 shows a plot of the mutual information and feature orientation distance measures applied along a horizontal search domain using a 9×9 sliding window. The maximum of the combined product is marked by the vertical red dashed line, and local maxima of the mutual information measure that are larger than the value of mutual information (used by itself) at this true maximum are marked by blue circles along the x-axis. The plot shows how the spatial similarity measure suppresses the many local maxima generated by the mutual information function.

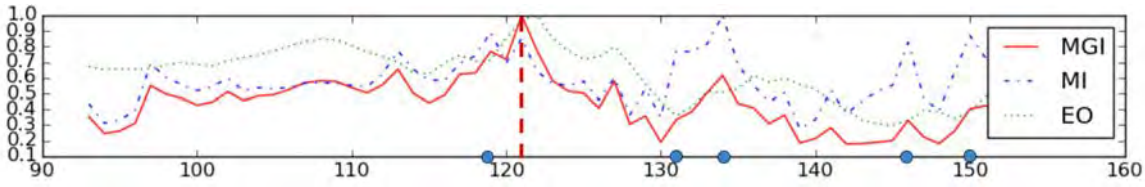


Figure 6.3: A line plot illustrating the search for correspondence using the MGI-PC similarity measure; the individual components are plotted to demonstrate how gradient information (EO) is used to suppress incorrect local maxima of mutual information (indicated with blue circles on the x-axis).

Figure 6.3 provides a qualitative example to demonstrate the purpose of spatial information in the combined measure. The investigation to demonstrate the performance of this combined measure and its application to computational stereo is described in the next section.

6.3 Hypothesis and objectives

Mutual information is a non-linear statistical method used extensively to register multi-spectral medical images. The first hypothesis put forward in this chapter is that mutual information can be used as a *local* similarity metric to match regions of thermal and visible spectrum images. A qualifier placed on this hypothesis is that the regions must be small enough to allow for accurate disparity estimation of points in the scene.

Mutual information can be applied to intensity maps or edge maps. The second hypothesis put forward is that alignment using mutual information with edge features, extracted by the phase congruency process, will improve the performance of the measure at larger scales. It is expected that the multitude of materials and structures of natural scenes will not provide consistent intensity value mappings and the joint entropy will remain dispersed. On the other hand, phase congruency was shown to increase visual similarity between the modalities in the structural EOH descriptor in Chapter 5, and mutual information using congruent edges is expected to provide clear maxima over a number of window scales.

To further increase performance, ambiguous maxima produced by the mutual information measure are enhanced by the incorporation of feature orientation from the reduced feature set of the phase congruency process. Feature orientation, indicated by the principal axis θ', was shown to be a repeatable means of comparing edges in

Section 6.6.

There are three objectives to the investigation presented in this chapter. Firstly, to demonstrate the performance of mutual information as a similarity metric for local regions of visible and thermal images. Secondly, to develop and present an enhanced measure using repeatable components of the phase congruency process. Finally, to show an improvement on the initial implementation of mutual information by demonstrating the performance of the enhanced similarity measure. An additional task, aimed at addressing the practical implementation and possible extensions of this method, is to show which scene elements are matched and at which scales (i.e. window sizes).

6.4 Methods

Local region matching in computational stereo involves taking a point from one image, called the reference image, and searching for it along the corresponding row in the other, called the query image, to find the point that maximises a cost function. This section presents the methods used to detect matchable points from the reference image and the approach to evaluation used to achieve the objectives set out for this chapter.

6.4.1 Entropy-based detection

Mutual information is built upon the Shannon entropy measure. Maximising mutual information of local regions involves a balance between maximising marginal entropies and minimising the joint entropy of the two regions [110]. Maxima of the marginal entropy map correspond to distinctive, information-rich regions which are stable and matchable; therefore, points are extracted from the marginal entropy map of the reference image to matched.

Interest points are extracted from the reference image by iteratively selecting the peak in the entropy map. A Non-maximal suppression (NMS) radial footprint is used to ensure adequate spatial distribution of the selected points. The surrounding region, within a radius of $N/3$ pixels of the peak, is set to zero so that subsequent points will not be detected within the suppression radius which scales with the size of the window

to control the overlap of adjacent regions. The motivation behind this is that if a large window is being used to estimate a coarse disparity map it does not make sense to attempt a fine-grained result by matching a large number of tightly clustered points.

6.4.2 Matching criteria and constraints

The input images are assumed to be rectified, so the search domain to match a point from the reference image is restricted to a horizontal disparity along the corresponding row in the query image. A maximum horizontal disparity is placed on the search domain; this constraint is standard practice in many applications and only requires that the minimum distance from the baseline to the scene is known. Additionally, without this limitation, the number of comparisons would make computation time unreasonably long with little added value.

A match is declared at the point that maximises the cost function if the marginal entropy of the query point is non-zero. Therefore, a match is only declared if the region in the query image contains matchable information. This qualifier on matching also decreases spurious responses caused by non-simultaneous phenomena.

6.4.3 Approach to evaluation

Evaluation is carried out by recording the ratio of true positives (i.e. correctly matched points) to the total number of declared correspondences. Two image pairs are available to be used in these automated tests and have a ground-truth disparity of ± 2 pixels; a correctly matched point is one that lies within this disparity (i.e. $-2 \leq d \leq 2$ pixels). The maximum disparity of the search domain is set to ± 40 pixels.

Evaluation is carried out at four scales; the square $N \times N$ block sizes are $N \in \{9, 15, 23, 27\}$ pixels with odd dimensions to ensure that the window is symmetric around the interest point. The number of points extracted from each image is dependent on the size of the non-maximal suppression window, which is defined as $N/3$ pixels. Note that no threshold was placed on the value of similarity (distance) of the declared matches.

The default parameters for the phase congruency process are shown in Table 4.1.

Mutual information requires that the edge map is quantised into discrete levels. The phase congruency edge map was heuristically quantised into $Q = 20$ levels to ensure that small congruency values were not filtered out (by quantising too coarsely) and to avoid increasing the sensitivity to noise (by quantising too finely). Although no explicit thresholding operations have been applied, this heuristic Q value does threshold the value of the edge magnitudes which is used as a significance measure for feature orientation.

The software libraries introduced in Section 4.3.4 were used in this investigation with the addition of pyentropy [111] to perform entropy and mutual information calculations.

6.5 Results

The evaluation of the proposed cost function is presented in this section. The results of the analysis are tabulated in Tables 6.1 and 6.2.

Comparison is carried out using the TPR expressed as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{6.5}$$

In order to compare different measures the number of correspondences is required to be roughly equal. The invariant representation provided by the phase congruency edge map has been shown to extract the stable information common to both spectra; however, the common edge information is sparsely distributed, resulting in large regions of zero-entropy. Figure 6.4 provides an example of the marginal entropy of one of the reference images used in this section.

The TPR, described in Eq. (6.5), is used to compare the performance of the distance measures presented in this section. The information measures Mutual information (MI) and Normalised mutual information (Y) are analysed separately and when combined, with multiplication, with the spatial Gradient information (GI). The product of mutual information (MI) and gradient information (GI) measures. (MGI). The product of normalised mutual information (Y) and gradient information (GI) (YGI). The information measures are used to compare regions of Intensity information (I) and Phase Congruency edge information (PC); this is expressed by appending the

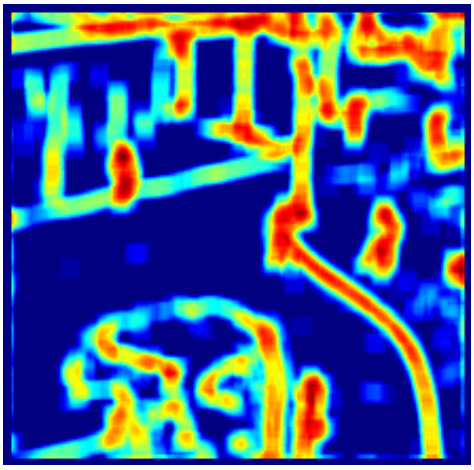


Figure 6.4: The marginal entropy H(X) at $N = 9$ of a phase congruency edge map. The dark blue regions show the regions of zero entropy that cannot be matched. (Image from the OSU database [44].)

measure with -I or -PC (e.g. MGI-PC is the product of the MI distance measure comparing regions of PC information and GI).

The results are presented in Tables 6.1 and 6.2. The analysis is carried out over four scales N (indicated in the first column if each table). A higher TPR indicates a better distance measure. However, as the scale N increases, an increasing difference in the number of regions detected when intensity information I or edge information PC was observed; therefore, the number of regions detected when either information source is used is indicated as two numbers separated by a forward slash in the second column (PC/I).

The results of the spatial information (GI) and the two mutual information measures are provided in Table 6.1. The results address the evaluation of mutual information as a local region descriptor over multiple scales with both intensity and congruency edge information.

Table 6.1: The TPR of mutual information measures MI and Y are analysed using phase congruency edges (PC) and intensity (I) information (the number of samples detected using each information source indicated in the PC/I column) over four scales N.

N	PC/I	MI-PC	MI-I	Y-PC	Y-I	GI
9	290/290	0.14	0.21	0.11	0.22	0.26
15	263/290	0.27	0.29	0.22	0.29	0.35
23	180/290	0.40	0.46	0.36	0.37	0.48
27	111/200	0.51	0.46	0.39	0.36	0.56

The proposed cost function combines mutual and spatial information; the two variations of mutual information are evaluated with different input information types. Spatial information is shown to improve the performance of the mutual information measures, particularly the phase congruency-based measures.

Table 6.2: The TPR of two combined measures, MGI and YGI, are analysed over four scales N with both phase congruency edges (PC) and intensity (I) information (the number of samples detected using each information source indicated in the PC/I column).

N	PC/I	MGI-PC	MGI-I	YGI-PC	YGI-I
9	290/290	0.24	0.29	0.23	0.25
15	263/290	0.42	0.40	0.39	0.39
23	180/290	0.58	0.57	0.54	0.53
27	111/200	0.68	0.62	0.64	0.59

Figures 6.5 and 6.6 show the distribution of true and false positives in the two images used in this investigation. In each case, the left image shows the correctly matched thermal regions superimposed on the visible spectrum image. On the right is the thermal image, with green and red crosses representing true and false positives respectively, with tails showing the displacement between the reference and query images.

(a) MGI-PC at $N = 15$ with TPR $= 0.40$.

(b) MGI-PC at $N = 23$ with TPR $= 0.55$.

Figure 6.5: Region mapping using the MGI-PC similarity measure. (Images from the OSU database [44].)

Figure 6.6c shows the regions that are detected and matched when the MGI metric is used with intensity information to provide a comparison of the distribution of matches when edge information, shown in Fig. 6.6a at the same scale, is used.

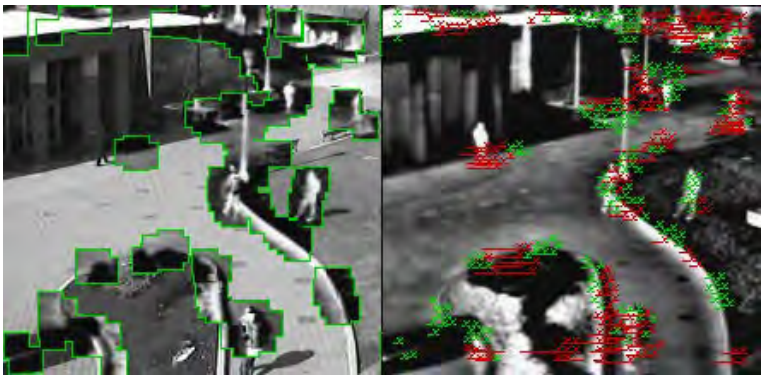
6.6 Discussion

The results presented in the previous section are primarily aimed at answering the initial hypotheses set out in the introduction of this chapter; however, the underlying goal of this discussion is to motivate further investigation into phase congruency and the proposed cost functions for multi-spectral image fusion. Keep in mind that performance over multiple scales is important, as coarse-to-fine searches and match propagation through scales is vital for accurate disparity estimation. The true positive rate (TPR, Eq. (4.18)) provides a general indication of the matching performance of the cost function over a number of scales.

6.6.1 Findings

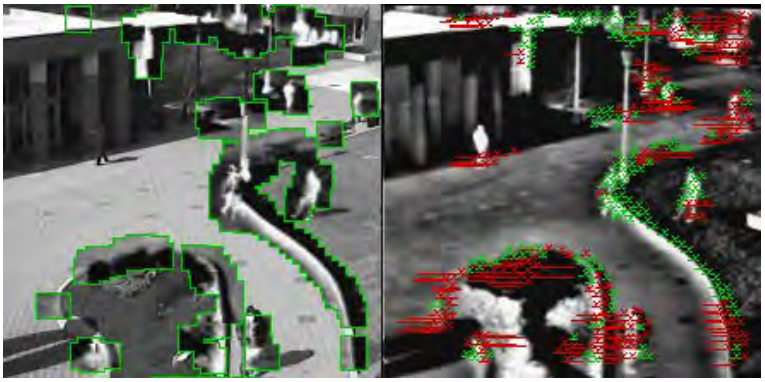
The two forms of mutual information presented in Section 6.1 are compared in Table 6.1 under data (i.e. edge map or intensity values) and scale (region size) variations. Both measures perform better when intensity information is used, with MI providing better performance than the normalised Y measure over all scales. It was hypothesised that, as the aperture N of the region increases, the inconsistent intensity mappings of the heterogeneous materials in natural scenes would cause the joint probability distribution to fail to develop the sharp peaks required to minimise the joint entropy value in the mutual information measure. The results in Table 6.1 show that mutual information is well suited to comparing and matching regions of visible and infrared intensity information over all the scales analysed, although the performance of the MI-I and Y-I measures does begin to drop off at the largest scale. As the intensity-based information measures are able to detect and match many more corresponding regions, despite the lower TPR, it is concluded that the use of the MI-I measure provides the best overall performance.

Combination of the mutual and spatial information measures is evaluated in Table 6.2. Although the MGI-PC measure has a slightly higher TPR, the MGI-I measure



(a) MGI-PC at $N = 15$ with TPR $= 0.50$.

(b) MGI-PC at $N = 23$ with TPR $= 0.70$.



(c) MGI-I at $N = 15$ with TPR $= 0.45$.

Figure 6.6: Region mapping using the MGI-PC and MGI-I similarity measures to demonstrate which regions are extracted and matched when congruent edges or intensity information is used. (Images from the OSU database [44].)

once again matches more points. The results presented in Table 6.2 show that the congruency-based spatial information GI improves the TPR performance of the MGI-PC and YGI-PC measures more than the intensity-based measures. Detection based on high entropy regions of the edge map selects regions with a large amount of edge information. As gradient information is only considered at points of congruency, regions that have more edges features will benefit more from spatial information. However, there is no guarantee that high entropy regions of intensity information will contain many edges (i.e. points of congruency), so these regions do not benefit from the incorporation of gradient information to the same degree.

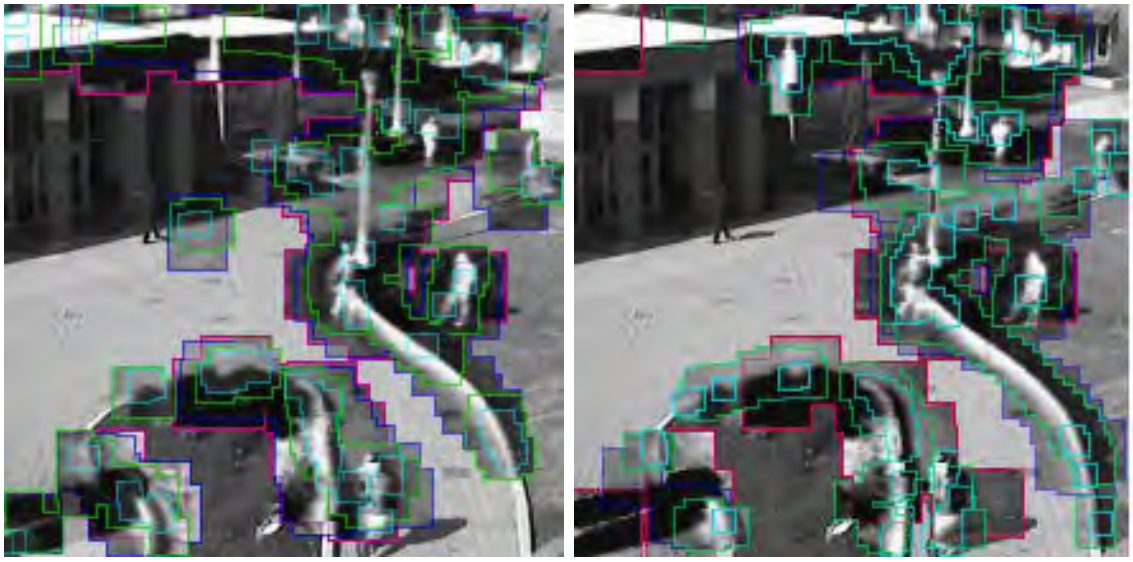
Although the spatial information, provided by GI, is best suited to the MI-PC measure, mutual information is also shown to perform well as a region descriptor for multi-spectral image intensity content. It is concluded that the combination of the MI-I and GI measures provided the best overall performance in terms of multi-scale matching and localisation accuracy at smaller scales. From a design point of view, it can be seen that the phase congruency and mutual information methods are well suited to work together, as both methods are designed to be adaptive to the underlying data and to operate with minimal heuristic input (e.g. thresholds, parameter tuning). Therefore it is expected that, despite the limited data available to this investigation, the results will generalise to different scenes and imaging devices.

6.6.2 Implications to fusion

The practical application and potential for further development of the MGI-PC is briefly discussed in this section. The goal of the qualitative analysis is to illustrate which structures are detected and matched, and to demonstrate the behaviour of the cost function over multiple scales.

Figures 6.7 and 6.8 show the correctly matched regions (using the same process shown in Fig. 6.5 and Fig. 6.6) superimposed with the largest scale on the bottom; each scale (the square window size) is outlined in a different colour.

Entropy-based detection identifies points containing distinctive information and attempts to find corresponding observations in the query image. The regions extracted therefore depend on the underlying image information.



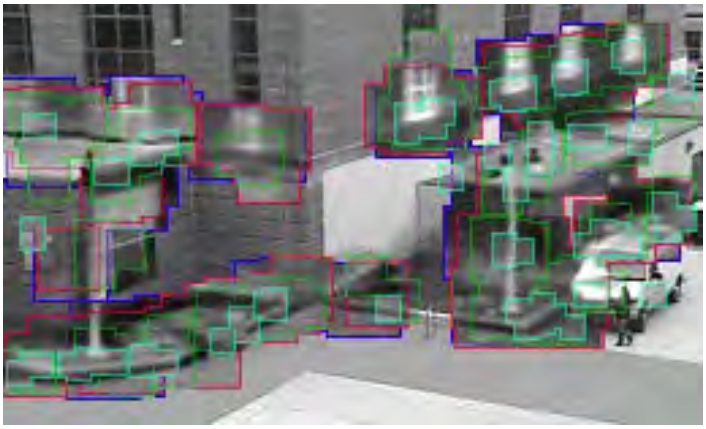
(a) **MGI-PC**. (b) **MGI-I**.

Figure 6.7: MGI-PC and MGI-I region mapping to show propagation of entropy, overlayed at multiple scales (square windows of $N \times N$ pixels where squares of size: $N = 9$ are light blue, $N = 15$ are light green, $N = 23$ are dark blue and $N = 27$ are red). (Images from the OSU database [44].)

Figures 6.7 and 6.8 show the structures that are matched when intensity information is used compared to when edge information is used. A noticeable difference between the two approaches is that the matches in the MGI-I images (Figs. 6.7b and 6.8b) appear more contour-like when overlapped over multiple scales. Based on this observation, it could be concluded that features detected using the intensity-based entropy function and matched with the MGI-I measure are able to localise structures over multiple scales more effectively than the PC-based detection and matching measures. In addition to this multi-scale matching stability, more matchable regions are detected when intensity information is used than when the PC edge maps are used; this can be seen in the *PC/I* column in Table 6.2 which shows that significantly more matchable features are found when detecting regions of high entropy in the intensity images, particularly at higher scales $N = \{23, 27\}$.

It is concluded that, despite the lower TPR to the MGI-PC function, the MGI-I function may provide more useful behaviour for fusion methods which utilise multi-scale structural matching information. However, more development in methods of multi-scale fusion with a larger image dataset is required to conclusively determine how well the MGI-I and MGI-PC methods are suited to multi-spectral image fusion.

(a) **MGI-PC**.



(b) **MGI-I**.

Figure 6.8: MGI-PC and MGI-I region-mapping to show propagation of entropy, overlayed at multiple scales (square windows of $N \times N$ pixels where squares of size: $N = 9$ are light blue, $N = 15$ are light green, $N = 23$ are dark blue and $N = 27$ are red) (Images from the OSU database [44].)

6.7 Summary

This chapter has presented the development and evaluation of multi-spectral cost functions for computational stereo. Variations of the region data and mutual information measure led to the conclusion that the regular mutual information metric performed well with the rich multi-spectral intensity information. It was also concluded that the incorporation of feature orientation information significantly improved performance at smaller scales. This is the second chapter that utilises the spectrum-invariant representation provided by phase congruency and further motivates its use with multi-spectral images, despite the computation time required by the current implementation. The next chapter provides a summary and brief discussion of the over all findings of this work.

Chapter 7

Conclusions and recommendations

The investigations carried out in this work aimed to identify effective methods that could be used to map multi-spectral observations from distinct viewpoints. The mapping performed by these methods synthesise a single enriched image in which each pixel has both colour and thermal information. Due to the application-specific requirements of fusion systems, information fusion was not addressed directly; instead, the methods required for fusion were the focus of this work.

The challenges and requirements to achieving this goal were initially clarified in Chapter 3. Parallax between the observations, captured from two viewpoints separated by a horizontal displacement, motivated the use of computational stereo methods and a region-based disparity mapping approach to overlaying the multi-spectral information. The focus was placed on the image alignment and correspondence matching stages of computational stereo. The motivation for this focus was based on the tasks' common requirement of a reliable and robust similarity measure to enable comparison of the disparate multi-spectral information.

Phase congruency, a frequency-domain analysis tool, was introduced and used in extracting repeatable and stable structural features. Chapter 4 presented an investigation with two objectives: to develop a stable representation of the two spectral modalities, and to isolate predictable features extracted in the phase congruency analysis process. Evaluation of the first of these objectives was deferred to its practical implementation in Chapters 5 and 6.

The second objective was accomplished using supervised learning with the SVM

model. The trained SVM was evaluated using the ROC to identify predictable relationships between phase congruency components extracted from the spectral modalities. Components were combined through multiplication and the cosine distance measure (which measures angular distance) based on knowledge drawn from an extended development of theory in Section 4.2. The investigation identified features with predictive value. The final pattern vector (seven components), which concluded the investigation process, provided the best AUC measure; however, the feature orientation distance of congruent edges showed comparable performance. Further analysis of the feature over the whole sample set showed its effectiveness as a repeatable similarity metric. As the feature orientation distance is a single number, it was easily integrated into the combined MI and GI (MGI) cost function, presented in Chapter 6, where its incorporation was shown to boost performance significantly.

Methods for automatic alignment of multi-spectral viewpoints were investigated in Chapter 5. The EOH descriptor was adapted to use the invariant representation (edge map) of the phase congruency process to replace the traditional gradient-based Canny edge detector. It was shown that conventional feature-based methods are unsuited for the variations in inter-spectrum image data. The Shi-Tomasi GF corner detector was found to be the most repeatable in the automated tests carried out, although it was observed that the phase congruency corner detector, while performing well in some cases, was prone to edge responses which introduced ambiguous matches (the aperture problem).

The EOH descriptor with the Shi-Tomasi GF corner detector (EOH-GF) performed significantly better than traditional feature detection and matching methods; however, its application is limited by its simple implementation which is sensitive to rotation, scaling and perspective distortion. Invariance to these factors is a fundamental requirement for alignment of uncalibrated cameras.

It is recommended that further development of the EOH descriptor is carried out to enable its application as a feature-point descriptor for unaligned images. However, until an effective inter-spectrum method for feature detection and description is developed, it is concluded that custom tools (such as those discussed in Section 3.2) are essential to accurately align the visible and thermal viewpoints.

Region-based correspondence search and matching of two aligned images was the second stage of computational stereo investigated in this work. Chapter 6 presented the development of a cost function incorporating mutual information and spatial features

extracted in the phase congruency process. It was found that mutual information performed similarly well with either multi-modal intensity or edge information; however, the use of edge information in detection and MI-PC matching saw greater benefits with the incorporation of spatial information. It is recommended that extensions to the cost function should focus on utilising this multi-scale performance; such methods include, for example, coarse-to-fine correspondence search, multi-scale match propagation and disparity map refinement.

This work has motivated the use of phase congruency as an essential step in adapting existing methods and developing new ones. The methods that were presented were largely unconstrained and can readily be improved for application-specific tasks. The significant difference in wavelength makes visible and thermal images the most difficult of the visible-infrared correspondence problems. The two components of the MGI-PC cost function (mutual information and phase congruency) are both applied without any heuristic thresholds; this important design objective means that the measure can, in theory, be applied to any multi-spectral pair. While the same could be said about the EOH feature descriptor, its application is limited in its current form.

Both the MGI-PC cost function and the EOH-GF feature descriptor demonstrate how phase congruency is well-suited to multi-spectral image analysis. While the computational cost of using the method makes it impractical for real-time applications in its current implementation, this computation time can be massively reduced as the algorithm is inherently parallel and suited to modern multi-threaded Central processing unit (CPU) or Graphics processing unit (GPU) architecture.

Frequency-domain analysis tools are essential when working with disparate multi-spectral images. Extracting information so that these modalities can be compared is a vital first step towards any framework for fusion. This work has demonstrated that phase congruency can perform this vital function and is a valuable tool for working with distinct visible and infrared spectra.

Bibliography

[1] D. Ziou, C. Armenakis, and D. Li, "A comparative analysis of image fusion methods," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, pp. 1391–1402, June 2005.

[2] J. B. Campo, "Multimodal Stereo from Thermal Infrared and Visible Spectrum," PhD thesis, Universitat Autònoma de Barcelona, Spain, 2012

[3] M. Govender, K. Chetty, and H. Bulcock, "A review of hyperspectral remote sensing and its application in vegetation and water resource studies," *Water SA*, vol. 33, no. 2, 2009.

[4] M. Moganti, F. Ercal, C. H. Dagli, and S. Tsunekawa, "Automatic PCB Inspection Algorithms: A Survey," *Computer Vision and Image Understanding*, vol. 63, pp. 287–313, Mar. 1996.

[5] H. Loh and M. Lu, "Printed circuit board inspection using image analysis," *IEEE Transactions on Industry Applications*, vol. 35, no. 2, pp. 426–432, 1999.

[6] A. Wong and W. Bishop, "Efficient least squares fusion of MRI and CT images using a phase congruency model," *Pattern Recognition Letters*, vol. 29, pp. 173–180, Feb. 2008.

[7] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE transactions on medical imaging*, vol. 16, pp. 187–98, Apr. 1997.

[8] J. P. W. Pluim, A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: A survey," *IEEE Transactions on Medical Imaging*, vol. 22, no. 8, pp. 986–1004, 2003.

[9] Z. Liu, D. S. Forsyth, and R. Laganière, "A feature-based metric for the quantitative evaluation of pixel-level image fusion," *Computer Vision and Image Understanding*, vol. 109, pp. 56–68, Jan. 2008.

[10] C. Pohl and J. L. Van Genderen, "Multisensor image fusion in remote sensing: Concepts, methods and applications," *International Journal of Remote Sensing*, vol. 19, no. 5, pp. 823–854, 1998.

[11] A. Toet, "Hierarchical image fusion," *Machine Vision and Applications*, vol. 3, no. 1, pp. 1–11, 1990.

[12] L. G. Brown, *A survey of image registration techniques*, PhD thesis, Columbia University, 1991.

[13] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[14] R. Luo and M. Kay, "A tutorial on multisensor integration and fusion," in *16th Annual Conference of IEEE Industrial Electronics Society (IECON)*, 1990.

[15] A. Rogalski, "Infrared detectors: Status and trends," *Progress in Quantum Electronics*, vol. 27, no. 2-3, 2003.

[16] A. Rogalski, "History of infrared detectors," *Opto-Electronics Review*, vol. 20, no. 3, pp. 279–308, 2012.

[17] M. Iqbal, *An introduction to solar radiation.* Elsevier, 1983.

[18] P. Ricaurte, C. Chilan, C. A. Aguilera-Carrasco, B. X. Vintimilla, and A. D. Sappa, "Feature point descriptors: Infrared and Visible Spectra," *Sensors 2014 (Switzerland)*, vol. 14, pp. 3690–3701, 2014.

[19] L. Schaul, C. Fredembach, and S. Süsstrunk, "Color image dehazing using the near-infrared," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 1629–1632, Citeseer, 2009.

[20] Y. M. Lu, C. Fredembach, M. Vetterli, and S. Susstrunk, "Designing color filter arrays for the joint capture of visible and near-infrared images," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 3797–3800, 2009.

[21] R. Fattal, "Single image dehazing," *ACM Transactions on Graphics*, vol. 27, no. 3, p. 1, 2008.

[22] D. Firmenichy, M. Brown, and S. Süsstrunk, "Multispectral interest points for RGB-NIR image registration," in *2011 18th IEEE International Conference on Image Processing (ICIP)*, pp. 181–184, 2011.

[23] J. Cronje and J. de Villiers, "A comparison of image features for registering LWIR and visual images," *23rd Annual Symposium of the Pattern Recognition Association of South Africa*, 2012.

[24] M. Bertozzi, a. Broggi, M. Felisa, G. Vezzoni, and M. Del Rose, "Low-level Pedestrian Detection by means of Visible and Far Infra-red Tetra-vision," *2006 IEEE Intelligent Vehicles Symposium*, pp. 231–236, 2006.

[25] J. Portmann, S. Lynen, M. Chli, and R. Siegwart, "People Detection and Tracking from Aerial Thermal Views," in *Autonomous Systems Lab, ETH Zurich*, 2014.

[26] T. T. Zin, H. Takahashi, T. Toriu, and H. Hama, "Fusion of Infrared and Visible Images for Robust Person Detection," in *Second International Conference on Innovative Computing, Information and Control ICICIC '07*, vol. 1, ch. 12, pp. 240–264, Osaka, Japan, 2007.

[27] S. Denman, T. Lamb, C. Fookes, V. Chandran, and S. Sridharan, "Multi-spectral fusion for surveillance systems," *Computers & Electrical Engineering*, vol. 36, pp. 643–663, July 2010.

[28] J. W. Davis and M. A. Keck, "A two-stage template approach to person detection in thermal imagery," in *Seventh IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 364–369, 2005.

[29] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893, 2005.

[30] T. I. Dhamecha, A. Nigam, R. Singh, and M. Vatsa, "Disguise detection and face recognition in visible and thermal spectrums," in *2013 International Conference on Biometrics (ICB)*, 2013.

[31] J. Bai, Y. Ma, J. Li, H. Li, Y. Fang, R. Wang, and H. Wang, "Good match exploration for thermal infrared face recognition based on YWF-SIFT with multi-scale fusion," *Infrared Physics & Technology*, vol. 67, pp. 91–97, Nov. 2014.

[32] R. Shoja Ghiass, O. Arandjelović, A. Bendada, and X. Maldague, "Infrared face recognition: A comprehensive review of methodologies and databases," *Pattern Recognition*, vol. 47, pp. 2807–2824, Sept. 2014.

[33] S. G. Kong, J. Heo, F. Boughorbel, Y. Zheng, B. R. Abidi, A. Koschan, M. Yi, and M. Abidi, "Multiscale Fusion of Visible and Thermal IR Images for Illumination-Invariant Face Recognition," *International Journal of Computer Vision*, vol. 71, pp. 215–233, June 2006.

[34] F. Barrera, F. Lumbreras, and A. D. Sappa, "Multispectral piecewise planar stereo using Manhattan-world assumption," *Pattern Recognition Letters*, vol. 34, pp. 52–61, Jan. 2013.

[35] C. Aguilera, F. Barrera, F. Lumbreras, A. D. Sappa, and R. Toledo, "Multispectral image feature points," *Sensors (Basel, Switzerland)*, vol. 12, pp. 12661–72, Jan. 2012.

[36] F. Barrera, F. Lumbreras, and A. D. Sappa, "Evaluation of Similarity Functions in Multimodal Stereo," pp. 320–329, 2012.

[37] F. Barrera, F. Lumbreras, and A. Sappa, "Multimodal template matching based on gradient and mutual information using scale-space," in *Proceedings of 2010 IEEE 17th International Conference on Image Processing (ICIP)*, (Hong Kong), pp. 2749–2752, 2010.

[38] C. Aguilera, F. Barrera, A. D. Sappa, and R. Toledo, "A Novel SIFT-Like-Based Approach for FIR-VS Images Registration," *11th International Conference on Qualitative InfraRed Thermography*, 2012.

[39] P. Kovesi, "What Are Log-Gabor Filters and Why Are They Good?," Internet: `http://www.csse.uwa.edu.au/~pk/Research/MatlabFns/PhaseCongruency/Docs/convexpl.html`, January 2015.

[40] P. Kovesi, "Phase congruency detects corners and edges," in *Proceedings of the VIIth Biennial Australian Pattern Recognition Society Conference (DICTA 2003)*, pp. 309–18, 2003.

[41] P. Kovesi, "Phase congruency: a low-level image invariant," *Psychological research*, vol. 64, pp. 136–148, Jan. 2000.

[42] P. Kovesi, "Edges Are Not Just Steps," in *The 5th Asian Conference on Computer Vision (ACCV2002)*, Melbourne, Australia, 2002.

[43] T. Mouats and N. Aouf, "Multimodal stereo correspondence based on phase congruency and edge histogram descriptor," *16th International Conference on Information Fusion*, pp. 1981–1987, Istanbul, Turkey, 2013.

[44] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Computer Vision and Image Understanding*, vol. 106, pp. 162–182, 2007.

[45] S. M. Seitz, J. Diebel, D. Scharstein, and R. Szeliski, "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms," 2001.

[46] M. Yaman and S. Kalkan, "Multimodal Stereo Vision Using Mutual Information with Adaptive Windowing," tech. rep., Computer Eng. Middle East Technical University, Ankara, Turkey, 2013.

[47] S. J. Krotosky and M. M. Trivedi, "Registering multimodal imagery with occluding objects using mutual information: application to stereo tracking of humans," in *Augmented Vision Perception in Infrared* (R. I. Hammoud, ed.), ch. 14, pp. 321—-347, London: Springer Science & Business Media, 2009.

[48] R. Szeliski, *Computer vision: algorithms and applications*. Springer, 2010.

[49] M. Mancuso and S. Battiato, "An Introduction to the Digital Still Camera Technology," *ST Journal of System Research*, vol. 2, no. 1, pp. 1–9, 2001.

[50] B. E. Bayer, "Color Imaging Array," 1976.

[51] S. Vidas, R. Lakemond, S. Denman, C. Fookes, S. Sridharan, and T. Wark, "A mask-based approach for the geometric calibration of thermal-infrared cameras," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, pp. 1625–1635, June 2012.

[52] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, pp. 247–259, May 2011.

[53] J. P. Heather and M. I. Smith, "Multimodal image registration with applications to image fusion," in *2005 8th International Conference on Information Fusion*, vol. 1, IEEE, 2005.

[54] M. Irani and P. Anandan, "Robust Multi-Sensor Image Alignment," in *1998 International Conference on Computer Vision*, pp. 1–19, 1998.

[55] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

[56] The MathWorks Inc., "Computer Vision System Toolbox," Internet:`http://www.vision.caltech.edu/bouguetj/calib_doc`, May 2014.

[57] Gary Bradski, "Camera calibration With OpenCV,", Internet:`http://www.opencv.org`, May 2014.

[58] A. J. Woods, T. Docherty, and R. Koch, "Image Distortions in Stereoscopic Video Systems," in *SPIE Symposium on Electronic Imaging: Stereoscopic Displays and Applications*, vol. 1915, San Jose, California, pp. 36–48, 1993.

[59] M. Z. Brown, D. Burschka, and G. D. Hager, "Advances in computational stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 8, pp. 993–1008, 2003.

[60] F. Tombari, S. Mattoccia, and L. D. Stefano, "Segmentation-based adaptive support for accurate stereo correspondence," *Advances in Image and Video Technology*, pp. 427–438, 2007.

[61] S. B. Kang, R. Szeliski, and J. C. J. Chai, "Handling occlusions in dense multi-view stereo," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, September 2001.

[62] S. Yoon, D. Min, and K. Sohn, "Fast dense stereo matching using adaptive window in hierarchical framework," *Advances in Visual Computing*, pp. 316–325, 2006.

[63] Nievergelt, Jürg and Preparata, Franco P., "Plane-sweep algorithms for intersecting geometric figures," *Communications of the ACM*, vol. 25, no. 10, pp. 739–747, 1982.

[64] J. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, 1986.

[65] M. C. Morrone and R. A. Owens, "Feature detection from local energy," *Pattern Recognition Letters*, vol. 6, no. 5, pp. 303–313, 1987.

[66] M. C. Morrone and D. C. Burr, "Feature Detection in Human Vision: A Phase-Dependent Energy Model," 1988.

[67] Z. Xiao and Z. Hou, "Phase based feature detector consistent with human visual system characteristics," *Pattern Recognition Letters*, vol. 25, pp. 1115–1121, July 2004.

[68] T. S. Lee, "Using 2D Gabor Wavelets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 10, pp. 1–13, 1996.

[69] Z. Wang and E. Simoncelli, "Local phase coherence and the perception of blur," in *Advances in neural information processing systems*, 2003.

[70] J. K. Kamarainen, V. Kyrki, and H. Kälviäinen, "Invariance properties of Gabor filter-based features - Overview and applications," *IEEE Transactions on Image Processing*, vol. 15, pp. 1088–1099, May 2006.

[71] S. Fischer, R. Redondo, G. Cristóbal, and I. D. O. Csic, "How to construct log-Gabor Filters?," tech. rep., Instituto de Optica (CSIC), Madrid, Spain, 2009.

[72] J. R. Movellan, "Tutorial on Gabor Filters," *Open Source Document*, 2002.

[73] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *Journal of the Optical Society of America A*, vol. 4, p. 2379, Dec. 1987.

[74] M. Mellor and M. Brady, "Phase mutual information as a similarity measure for registration," *Medical Image Analysis*, vol. 9, pp. 330–343, Aug. 2005.

[75] A. Ng, "Support Vector Machines.", *CS229 (Stanford) lecture note series*, 2000.

[76] G. M. Foody and J. Mathur, "A relative evaluation ofmulticlass image classification by support vector machines.," *IEEE Trans. Geosci. Rem. Sens*, vol. 1343, pp. 1–9, 2004.

[77] P. D. Kovesi, "MATLAB and Octave Functions for Computer Vision and Image Processing," Internet: `http://www.peterkovesi.com/matlabfns/index.html`, May 2014.

[78] S. Van Der Walt, S. C. Colbert, and G. Varoquaux, "The NumPy array: a structure for efficient numerical computation," *Computing in Science & Engineering*, vol. 13, no. 2, pp. 22–30, 2011.

[79] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[80] T. Lindeberg, "Feature Detection with Automatic Scale Selection," *International Journal of Computer Vision*, vol. 30, no. 2, pp. 79 – 116, 1998.

[81] C. Harris and M. Stephens, "A Combined Corner and Edge Detector," *Procedings of the Alvey Vision Conference 1988*, pp. 23.1–23.6, 1988.

[82] H. P. Moravec, "Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover," Technical Report, DTIC Document, 1980.

[83] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," *Computer Vision–ECCV 2006*, Springer, pp. 430–443, 2006.

[84] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," *2011 International Conference on Computer Vision*, pp. 2564–2571, Nov. 2011.

[85] J. Cronje, "BFROST: Binary Features from Robust Orientation Segment Tests accelerated on the GPU," *22nd Annual Symposium of the Pattern Recognition Association of South Africa*, 2011.

[86] J. Heinly, E. Dunn, and J. M. Frahm, "Comparative evaluation of binary features," *Computer Vision–ECCV 2012*, Springer, pp. 759–773, 2012.

[87] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary Robust invariant scalable keypoints," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2548–2555, 2011.

[88] E. Mair, G. D. Hager, D. Burschka, M. Suppa, and G. Hirzinger, "Adaptive and generic corner detection based on the accelerated segment test," *Computer Vision–ECCV 2010*, Springer, pp. 183–196, 2010.

[89] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, Nov. 2004.

[90] H. Bay, A. Ess, T. Tuytelaars, and L. van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[91] J. Han, E. Pauwels, and P. D. Zeeuw, "Visible and infrared image registration employing line-based geometric analysis," *Computational Intelligence for Multimedia Understanding*, vol. 7252, pp. 114–125, 2012.

[92] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary robust independent elementary features," *Computer Vision–ECCV 2010*, 2010.

[93] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast retina keypoint," in *Proceedings of the IEEE Computer Society conference on Computer Vision and Pattern Recognition*, 2012.

[94] M. Teke, M. F. Vural, A. Temizel, and Y. Yardmc, "High-resolution multispectral satellite image matching using scale invariant feature transform and speeded up

robust features," *Journal of Applied Remote Sensing*, vol. 5, no. 1, pp. 053553-1–053553-9, 2011.

[95] A. Sedaghat, M. Mokhtarzade, and H. Ebadi, "Uniform robust scale-invariant feature matching for optical remote sensing images," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 11, pp. 4516–4527, 2011.

[96] D. K. Park, Y. S. Jeon, and C. S. Won, "Efficient use of local edge histogram descriptor," *Proceedings of the 2000 ACM workshops on Multimedia*, pp. 51–54, 2000.

[97] T. Sikora, "The MPEG-7 Visual Standard for Content Description–An Overview," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 696–702, 2001.

[98] B. Alefs, G. Eschemann, H. Ramoser, and C. Beleznai, "Road Sign Detection from Edge Orientation Histograms," *2007 IEEE Intelligent Vehicles Symposium*, pp. 993–998, June 2007.

[99] W. Freeman and M. Roth, "Orientation histograms for hand gesture recognition," *International Workshop on Automatic Face and Gesture Recognition*, vol. 12, pp. 296–301, 1995.

[100] S. Sonn, G.-A. Bilodeau, and P. Galinier, "Fast and Accurate Registration of Visible and Infrared Videos," *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 308–313, June 2013.

[101] G. Egnal, "Mutual Information as a Stereo Correspondence Measure," Technical Report (CIS), Department of Computer Science, University of Pennsylvania, January 2000.

[102] A. Torabi, M. Najafianrazavi, and G. A. Bilodeau, "A comparative evaluation of multimodal dense stereo correspondence measures," in *Proceedings of IEEE International Symposium on Robotic and Sensors Environments*, pp. 143–148, 2011.

[103] A. Torabi and G. A. Bilodeau, "Local self-similarity as a dense stereo correspondence measure for themal-visible video registration," Technical Report, Ecole Polytechnique de Montreal, Montreal, Canada, 2011.

[104] P. Viola and W. Wells, "Alignment by maximization of mutual information," in *Proceedings of IEEE International Conference on Computer Vision*, vol. 24, no. 2, pp. 137–154, 1997.

[105] R. P. Woods, J. C. Mazziotta, and S. R. Cherry, "MRI-PET registration with automated algorithm," *Journal of computer assisted tomography*, vol. 17, no. 4, pp. 536–546, 1993.

[106] R. P. Woods, S. R. Cherry, and J. C. Mazziotta, "Rapid automated algorithm for aligning and reslicing PET images.," *Journal of computer assisted tomography*, vol. 16, no. 4, pp. 620–633, 1992.

[107] S. Suri, P. Schwind, P. Reinartz, and J. Uhl, "Combining Mutual Information and Scale Invariant Feature Transform for Fast and Robust Multisensor SAR Image Registration," *75th Annual ASPRS Conference*, no. I, 2009.

[108] E. G. Learned-Miller, "Entropy and Mutual Information," Department of Computer Science, University of Massachusetts, Amherst, 2013.

[109] T. M. Cover and J. A. Thomas, "Entropy , Relative Entropy and Mutual Information," in *Elements of Information Theory*, ch. 2, pp. 12–49, John Wiley & Sons, Inc., 1 ed., 1991.

[110] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Image registration by maximization of combined mutual information and gradient information," *IEEE Transactions on Medical Imaging*, vol. 19, no. 8, pp. 809–814, 2000.

[111] R. A. A. Ince, R. S. Petersen, D. C. Swan, and S. Panzeri, "Python for information theoretic analysis of neural data.," *Frontiers in Neuroinformatics*, vol. 3, p. 4, Feb. 2009.