Universitat Autònoma de Barcelona

Local feature description in cross-spectral imagery

Cristhian Alejandro Aguilera Carrasco

ADVERTIMENT. L'accés als continguts d'aquesta tesi doctoral i la seva utilització ha de respectar els drets de la persona autora. Pot ser utilitzada per a consulta o estudi personal, així com en activitats o materials d'investigació i docència en els termes establerts a l'art. 32 del Text Refós de la Llei de Propietat Intel·lectual (RDL 1/1996). Per altres utilitzacions es requereix l'autorització prèvia i expressa de la persona autora. En qualsevol cas, en la utilització dels seus continguts caldrà indicar de forma clara el nom i cognoms de la persona autora i el títol de la tesi doctoral. No s'autoritza la seva reproducció o altres formes d'explotació efectuades amb finalitats de lucre ni la seva comunicació pública des d'un lloc aliè al servei TDX. Tampoc s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant als continguts de la tesi com als seus resums i índexs.

ADVERTENCIA. El acceso a los contenidos de esta tesis doctoral y su utilización debe respetar los derechos de la persona autora. Puede ser utilizada para consulta o estudio personal, así como en actividades o materiales de investigación y docencia en los términos establecidos en el art. 32 del Texto Refundido de la Ley de Propiedad Intelectual (RDL 1/1996). Para otros usos se requiere la autorización previa y expresa de la persona autora. En cualquier caso, en la utilización de sus contenidos se deberá indicar de forma clara el nombre y apellidos de la persona autora y el título de la tesis doctoral. No se autoriza su reproducción u otras formas de explotación efectuadas con fines lucrativos ni su comunicación pública desde un sitio ajeno al servicio TDR. Tampoco se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al contenido de la tesis como a sus resúmenes e índices.

WARNING. The access to the contents of this doctoral thesis and its use must respect the rights of the author. It can be used for reference or private study, as well as research and learning activities or materials in the terms established by the 32nd article of the Spanish Consolidated Copyright Act (RDL 1/1996). Express and previous authorization of the author is required for any other uses. In any case, when using its content, full name of the author and title of the thesis must be clearly indicated. Reproduction or other forms of for profit use or public communication from outside TDX service is not allowed. Presentation of its content in a window or frame external to TDX (framing) is not authorized either. These rights affect both the content of the thesis and its abstracts and indexes.

Local feature description in cross-spectral imagery

A dissertation submitted by **Cristhian Alejandro Aguilera Carrasco** at Universitat Autònoma de Barcelona to fulfil the degree of **Doctor of Philosophy**.

Bellaterra, September 4, 2017

Director	**Dr. Angel D. Sappa** Centro de Visión por Computador
Thesis committee	**Arturo de la Escalera Hueso** Dpto. Ingeniería de Sistemas y Automática Universidad Carlos III de Madrid **Domenec Savi Puig Valls** Dpto. Ingeniería Informática y Matemáticas Universitat Rovira i Virgili **Carlos Alejandro Parraga** Dept. Ciencias de la Computación & Centro de Visión por Computador Universitat Autònoma de Barcelona

This document was typeset by the author using LaTeX 2_ε.

The research described in this book was carried out at the Centre de Visió per Computador, Universitat Autònoma de Barcelona.

Copyright © 2017 by Cristhian Alejandro Aguilera Carrasco. All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the author.

ISBN: XXX-XX-XXXXXX-X-X

Printed by Ediciones Gráficas Rey, S.L.

Acknowledgements

After four years of studying in Barcelona, I feel grateful to many people and life itself.

I would like to express my sincere admiration to the CVC staff, especially to Claire and Montse, two of the most efficient and amazing administrative that I have ever seen in my life.

To all the Ph.D. students that I meet during this trip. In particular, thanks to Prassana for being such a good friend. To Pau for having an enormous patience teaching me how to manage ELCVIA, and to Arash and Anguelos, for always having interesting discussions non-related to work.

To my advisors, Angel and Ricardo. They always had the best predisposition to help me when I needed, even in my personal affairs.

Lastly, but not less important, to my Family. To my parents and my brother to support me all these years. To Angela, for being the mother of the most important person in my life. And to Roser, the cutest daughter that I could ever ask for. So cute, that with one smile she can make my day.

Abstract

Over the last few years, the number of consumer computer vision applications has increased dramatically. Today, computer vision solutions can be found in video game consoles, smartphone applications, driving assistance—just to name a few. Ideally, we require the performance of those applications, particularly those that are safety-critical to remain constant under any external environment factors, such as changes in illumination or weather conditions. However, this is not always possible or very difficult to obtain by only using visible imagery, due to the inherent limitations of the images from that spectral band. For that reason, the use of images from different or multiple spectral bands is becoming more appealing.

The aforementioned possible advantages of using images from multiples spectral bands on various vision applications make multi-spectral image processing a relevant topic for research and development. Like in visible image processing, multi-spectral image processing needs tools and algorithms to handle information from various spectral bands. Furthermore, traditional tools such as local feature detection, which is the basis of many vision tasks such as visual odometry, image registration, or structure from motion, must be adjusted or reformulated to operate under new conditions. Traditional feature detection, description, and matching methods tend to underperform in multi-spectral settings, in comparison to mono-spectral settings, due to the natural differences between each spectral band.

The work in this thesis is focused on the local feature description problem when cross-spectral images are considered. In this context, this dissertation has three main contributions. Firstly, the work starts by proposing the usage of a combination of frequency and spatial information, in a multi-scale scheme, as feature description. Evaluations of this proposal, based on classical hand-made feature descriptors, and comparisons with state of the art cross-spectral approaches help to find and understand limitations of such strategy. Secondly, different convolutional neural network (CNN) based architectures are evaluated when used to describe cross-spectral image patches. Results showed that CNN-based methods, designed to work with visible monocular images, could be successfully applied to the description of images from two different spectral bands, with just minor modifications. In this framework, a novel CNN-based network model, specifically intended to describe image patches from two different spectral bands, is proposed. This network, referred to as Q-Net, outperforms state of the art in the cross-spectral domain, including both previous hand-made solutions as well as L2 CNN-based architectures. The third contribution of this dissertation is in the cross-spectral feature description application domain. The multispectral

odometry problem is tackled showing a real application of cross-spectral descriptors In addition to the three main contributions mentioned above, in this dissertation, two different multi-spectral datasets are generated and shared with the community to be used as benchmarks for further studies.

Resumen

En los últimos años, el número de aplicaciones de consumo basadas en visión por computadora han incrementado drásticamente. Actualmente, soluciones basadas en visión por computadora pueden ser encontradas en video juegos, aplicaciones móviles y en automóviles, por nombrar algunas. Idealmente, el desempeño de estas aplicaciones debiera ser igual ante cualquier factor externo, como cambios en la iluminación o del clima. Sin embargo, esto no es siempre posible utilizando sólo información del espectro visible, debido a las limitaciones inherentes de las imágenes de esta banda espectral. Razón por la cual, el uso de imágenes de diferentes bandas espectrales se está volviendo más común.

Las posibilidades que ofrece el uso de imágenes de diferentes espectros, hacen que su estudio sea un tema relevante de investigación y desarrollo. Al igual que, en el caso monocular, el procesamiento de imágenes multiespectrales necesita de algoritmos que puedan manejar su información. Herramientas tradicionales como descriptores locales de características, que son la base de varias técnicas de visión por computadora, deben ser ajustadas para operar en estas nuevas condiciones. Métodos tradicionales de detección, descripción y correspondencia, suelen tener un desempeño limitado en entornos multispectrales, al compararlos con su desempeño en el caso monocular visible. Esto se debe, principalmente a las diferencias naturales que existen entre las diferentes bandas espectrales, no consideradas en su diseño.

En esta tesis, nos enfocamos en el problema de la descripción de características locales de imágenes provenientes de diferentes bandas espectrales. En este contexto, el trabajo que se presenta contiene tres grandes contribuciones. En una primera instancia, propone el uso combinado de información frecuencial y espacial para la descripción de imágenes. Luego, realiza un estudio de diferentes técnicas basadas en redes convolucionales para describir imágenes provenientes de diferentes bandas espectrales. Los resultados muestran que este tipo de técnicas sobrepasan los resultados obtenidos por descriptores clásicos. En esta línea, presentamos una nueva red llamada Q-Net, que mejora el estado del arte en descriptores multiespectrales basados en redes convolucionales. La tercera contribución es una propuesta para el uso de estos nuevos descriptores en una aplicación de visión por computadora. En concreto, enfrentamos el problema de odometria visual, utilizando imágenes de diferentes espectros. Finalmente, dos conjuntos de datos fueron generados y compartidos con la comunidad científica en el desarrollo de esta tesis, que esperamos sean utilizados en estudios por otros investigadores.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The usage of images from different spectral bands opens new opportunities to devise novel solutions to traditional vision tasks. For example, due to the inherent limitations of images from the visible spectrum, the recognition of faces in low-illumination scenarios, like night-time surveillance, becomes a challenging or almost impossible task. In contrast, the recognition of faces using Near-Infrared (NIR) imagery is robust against different illumination changes, making it a more suitable choice for night-time surveillance [31]. Furthermore, solutions can involve the usage of two or more cameras from different spectral bands at the same time to complement the limitations of each one apart.

Images from a particular spectral band can provide several benefits and can also have several drawbacks for a given vision task. Therefore, the decision on how many cameras to use and from which spectral band is an application oriented task. For example, Long-Wave infrared images (LWIR), also referred to as thermal images, are practical to segment people from the background; especially at night when in general the environment temperature is lower than the human body. On the contrary, thermal images are not as practical as visual cameras to recognize people faces. Hence, the simultaneous use of LWIR and visible cameras is almost a standard in high-end video surveillance systems.

The aforementioned possible advantages of using images from multiples spectral bands on various vision applications make multi-spectral image processing a relevant topic for research and development. Consequently, many products that use cameras from different spectral bands have starting to become more common in the last few years. For instance, the Microsoft Kinect2 is a widely sold product that uses a visible camera along with a NIR camera for detecting and capturing the 3D motion of players, replacing in that way traditional game controllers. Other examples are the HeatWave system [55], which makes use of a visible, a NIR, and a LWIR cameras to create a 3D thermal image of a building, and the FLIR ONE system that provides an enhanced thermal vision for mobile devices using a LWIR and a visible camera.

Like in visible image processing, multi-spectral image processing needs tools and algorithms to handle information from various spectral bands. Furthermore, traditional tools such as local feature detection, which is fundamental in vision tasks such as visual odometry, image registration, and structure from motion, just to mention a few, must be adjusted or reformulated to operate under new conditions. Traditional feature detection, description, and matching methods tend to underperform in multi-spectral settings, in comparison to mono-spectral settings, due to the natural differences between each spectral band. Figure 1.1 shows an example of feature matching between a visible and a LWIR image using SIFT [33], where most of the feature matches are wrong.

Figure 1.1: The figure shows the result of performing feature matching between a visible and an LWIR image using SIFT. Visible image at the left and LWIR at the right.

Although several methods exist in the literature to find correspondences between two or more images, not many are useful in multi-spectral scenarios. For that reason, new contributions have starting to appear in the literature to tackle the multi-spectral matching problem (e.g., [5, 32, 48]). Unfortunately, their performance is far away from the one obtained when images from the same sensor and setup are considered. Therefore, new lines of research are opening, not only regarding the feature matching process but additionally in the possible applications.

1.1 Research objectives

The primary objective of this thesis is to address the problem of visible-infrared image description, i.e., cross-spectral description, and its potential applications to computer vision tasks. In particular, we are interested in:

- Improving the feature matching performance of current methods used to find correspondences between images from the visible and infrared spectral bands, through the proposal of new cross-spectral feature descriptor algorithms. The work is mainly focused on two cross-spectral cases, namely, VIS-NIR and VIS-LWIR.

- Exploring different uses of cross-spectral imagery in computer vision applications, especially those that involve cross-spectral feature description in one of their steps.
- Acquiring and sharing new multi-spectral datasets to be used as benchmarks for further research in this field. Our intention is to help other researchers to have access to multi-spectral datasets that in many cases are of difficult access due to the cost of the cameras.

1.2 Contributions

The contributions of this thesis can be summarized as follows:

- We acquired and shared two different multi-spectral datasets. One to evaluate VIS-LWIR cross-spectral descriptors and another to evaluate VIS-LWIR cross-spectral visual odometry solutions.
- We modified an existing VIS-NIR dataset for scene category recognition, to be used to train and evaluate CNN-based cross-spectral descriptors.
- We propose, evaluate and validate two different cross-spectral feature descriptor algorithms. One, using classic computer vision approaches and the other using a new convolutional neural network architecture.
- Finally, we proposed and evaluated two possible real applications that utilize cross-spectral feature descriptors in one of their steps.

1.3 Outline

The thesis is organized as follows. In Chapter 2, we describe the most significant similarities and differences between images from the visible and infrared spectral bands. Additionally, we discuss previous work on cross-spectral feature description, CNN-based description, and cross-spectral applications. In Chapter 3, we introduce two new visible-infrared benchmarks that are used in following chapters to validate our different proposals. In Chapter 4, we propose a cross-spectral feature descriptor based on the use of Log-Gabor filters. In Chapter 5, we explore different CNN-based solutions to describe visible and infrared image patches, and in Chapter 6, we propose a new CNN-based architecture to train cross-spectral feature descriptors that can be used as a drop-in replacement of classical feature descriptors, such as SIFT or SURF. In Chapter 7, we explore and evaluate two applications of cross-spectral descriptors; one for registering images as a previous step to image fusion and the other to determine the location of a car through visual odometry. Finally, Chapter 8 summarizes the result of our work and provides insights for further research in the area.

Chapter 2

Background and related work

The goal of this chapter is twofold. Firstly, to provide a basic understanding of the similarities and differences between images from the visible spectrum and images from the NIR and LWIR spectral bands, which are the two cross-spectral bands used to evaluate the different proposal in the current work. Secondly, to review the most relevant methods proposed in the literature to describe local features in images from two different spectral bands, and current CNN-based trends used to describe image patches.

This chapter will help to understand the work and contributions presented in following chapters.

2.1 Background

Modern cameras can capture electromagnetic radiations from multiple ranges of the electromagnetic spectrum (see Figure 2.1). Typical consumer cameras are sensitive to visible light, wavelengths between 0.3 and 0.7 μm, which is also visible to our eyes. Others are sensitive to other ranges of wavelengths, non-visible to our eyes, such as the infrared range between 0.7 and 14 μm, and the ultraviolet range between 10 and 400 nm; to name a few. In particular, we are interested in cameras from two subbands of the infrared spectrum, the NIR and LWIR sub-bands described in the next two sections. These two subbands have been selected for the following reasons. The NIR spectral band has been selected due to the low cost of the cameras needed to acquire images, while the LWIR spectral band has been selected for the specific characteristics of images acquired at this spectral band—thermal radiation.

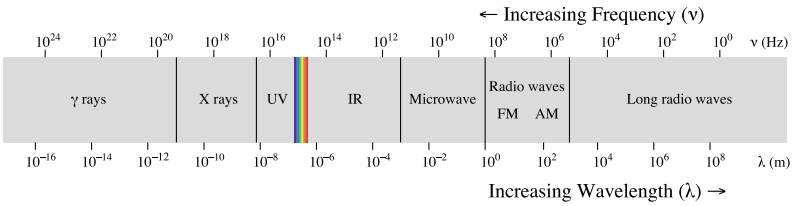


Figure 2.1: The electromagnetic spectrum.

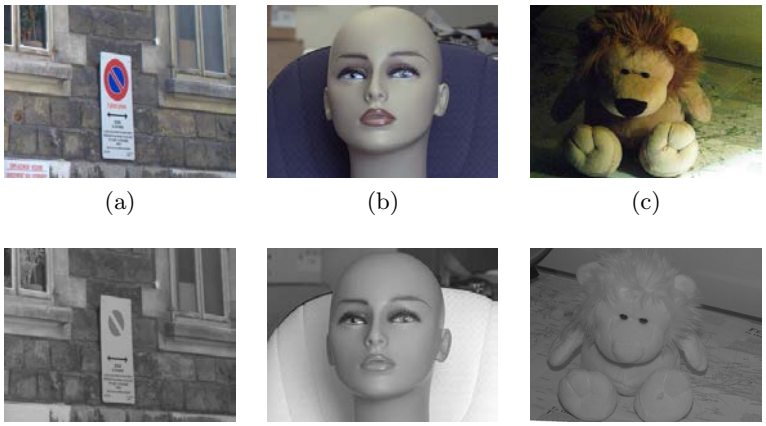


Figure 2.2: VIS-NIR image pairs; top images are from the visible spectrum and bottom images from the near-infrared spectrum.

2.1.1 NIR imaging

The near-infrared spectral band (0.7-1.4 μm) is one of the five sub-bands of the infrared spectrum. It is immediately adjacent to the visible portion of the electromagnetic spectrum, and as a consequence, camera sensors from both types of cameras are sensitive to some piece of the other spectral band. For example, to prevent unnatural looking images, visible cameras sensors need to include an *infrared cut-off filter* to block NIR wavelengths; visible camera sensors are often sensitive up to 1 μm approximately.

There is a wide number of uses for NIR cameras in consumer and industrial vision applications. On the consumer side, we can list a few examples, such as baby wireless monitors, video game consoles, and home security cameras, that use NIR cameras to view in low-lighting scenarios. These types of camera can see in the dark through active NIR illumination, which is not visible to human eyes, making it a perfect solution for low-light vision tasks that require not disturbing people attention. On the industrial side, these cameras are used to detect bruises in apple [24], to detect fatigue in driver during long trips [19], and to classify different types of breast cancer [56];

amongst many others.

Images from the visible and NIR spectral bands share several visual similarities. In Figure 2.2 three pairs of VIS-NIR images are presented. It can be appreciated that images from both spectral bands are visually similar but with notable differences. For example, red visible regions disappear in NIR images (see Figure 2.2d), the direction of gradients may be different as in Figure 2.2b in contrast with Figure 2.2e, and NIR images are more likely to be robust to different illumination settings as in Figure 2.2f in comparison with Figure 2.2c.

Therefore the challenge is in being able to develop a feature descriptor that is robust against the above-discussed differences.

2.1.2 LWIR imaging

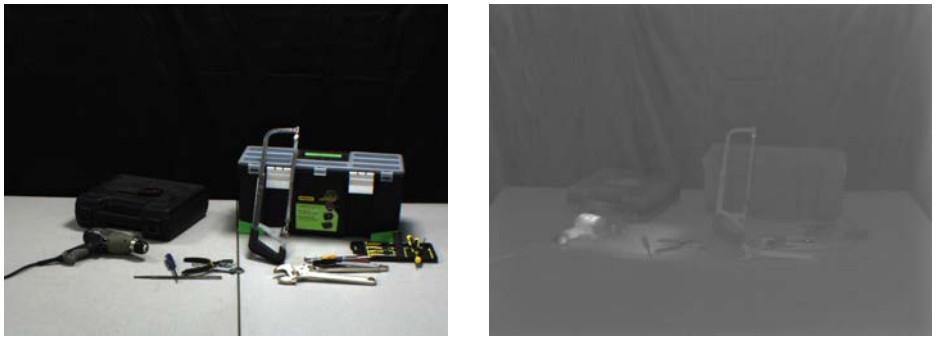
LWIR cameras, also called thermal cameras, are sensitive to infrared electromagnetic radiation wavelengths between 7 and 14 μm. Historically, were developed and used for military purposes, but now, thanks to recent technological advances and the reduction in the manufacturing cost, are available to all type of non-military applications. For instance, thermal cameras are used to detect and estimate wild animal populations [11], to detect pedestrians in advanced driver-assistance systems [22], to detect gas leaks [30], and to prevent and monitor injuries in sports [13]. For more examples, we refer the readers to the work of Gade & Moeslund in 2014 *Thermal cameras and applications: a survey* [18].

Thermal cameras capture the infrared radiation emitted by any object with a temperature above the absolute zero. In other words, thermal cameras map the temperature of all the objects in a scene to an image. Therefore, thermal images do not require visible light to work, and pixels' intensities are related to the temperature of the objects rather than lighting conditions. As a consequence, thermal cameras are less sensitive to illumination problems, such as pitch black and shadow areas, in contrast to visible cameras.

Images from thermal cameras have a significant visual difference with images taken from visible cameras (see Figure 2.3 and Figure 2.4). In contrast to visible images, thermal images do not have color and high-frequency information tend to be lost; the former since thermal cameras captures heat radiation from bodies rather than light, and the latter due to the homogeneity in temperature that exists in many objects, especially in human-made structures. Moreover, thermal images often can display *ghost objects*, which are objects that once were on the scene but were no longer there when the image was taken. This effect occurs since temperature changes in objects happen relative slow. Therefore, the temperature of an object not only depends on the current state of the object but additionally in the recent history.

Although, thermal and visible images may have many visual differences as explained above, also may share similarities. Studies like [38] suggest that there is a strong correlation between object boundaries in images from both spectral bands. In other words, even though most of the texture may be lost in thermal images, the shape

of the objects is preserved. Moreover, shape similarities are more prone to occur in environments with high-temperature variations, such as outdoor environments. On the contrary, in indoor environments, the shape of objects is less preserved, due to the low variations of temperature in a controlled environment. For example, Figure 2.3b is much less detailed than Figure 2.4b.



(a) Visible image (b) Thermal image

Figure 2.3: VIS-LWIR: indoor images from the same scene taken with cameras from different spectral bands.

(a) Visible image (b) Thermal image

Figure 2.4: VIS-NIR: outdoor images from the same scene taken with cameras from different spectral bands.

Therefore, the challenge to find correspondences between images from the visible spectrum and images from the LWIR spectrum lies in the capability to be able to design a feature descriptor that gives more weight to object boundaries, visible in both types of images, rather than color and texture, which are only available in visible images.

2.2 Related Work

In this section, we review the most relevant works in the literature that deal with the problem of cross-spectral feature description. Additionally, we also review the state-of-the-art of CNN-based descriptors that will be fundamental to understand our proposal in Chapter 6. Lastly, we review existing applications of cross-spectral (visible and infrared spectra) imagery.

2.2.1 Cross-spectral description

The description of image patches from two different spectral bands is a complex and challenging task. Classical feature descriptors, such as SIFT [33], SURF [6] and BRIEF [10], do not take into account the nonlinear intensity variations that may exist. Moreover, the performance of those algorithms tends to decrease when applied to images beyond the visual spectrum [44]. Early efforts to describe images from different spectral bands focused on modifying gradient-based descriptors to work between $[0, \pi]$ instead of $[0, 2\pi]$. This adjustment, reduce the unwanted effects of changes in gradient directions between images that can be typically found in these type of scenarios; as discussed in Section 2.1.1. For example, Firmenichy et al. in 2011 [14], changed the way the SIFT descriptor is computed. Originally, SIFT is computed as in Figure 2.5. The image patch initially is divided into sixteen sub-regions, and gradient intensities directions are computed for all pixels in the image; directions have only eight possible values equally divided between $[0, 2\pi]$. Then, a histogram of the gradients' direction is computed, bins per sub-region, with a total of 128 bins (16x8). Instead of this approach, Firmenichy et al. defined only four directions between $[0, \pi]$, eliminating complementary angles, and finally obtaining a descriptor of size 64 (4x16) (see Figure 2.5). This algorithm, named GSIFT, drastically improves SIFT performance in multi-spectral scenarios. Another example is the work of Pinggera et al. in 2012 [43], where the authors modified the HOG descriptor to compute gradient directions between $[0, \pi]$ instead of $[0, 2\pi]$, to calculate the stereo disparity between a visible and an LWIR camera. In their experiments, they discovered that such a strategy performed better than traditional multi-spectral techniques such as mutual information and local self-similarity.

Other works follow the observations of Morris et al. [38]. i.e., giving more importance to shape and contours rather than to texture. In their study, concerning the joint statistics of visible and thermal images, the authors found a strong correlation between object boundaries of images from both spectra, in other words, although texture information may be lost, the shape of objects remain similar between images from both spectral bands. In this context, Aguilera et al. [1] decided to describe multi-spectral image patches using a local version of the global Edge-Oriented-Histogram (EHD) descriptor [35]. The proposed approach consisted of using a histogram of contours' orientation in the surroundings of each interest point. Contours' orientation are obtained using five 3x3 Canny filters, where four filters are used to indicate an orientation between $[0, \pi]$, and the last filter to indicate no-direction. After the com-

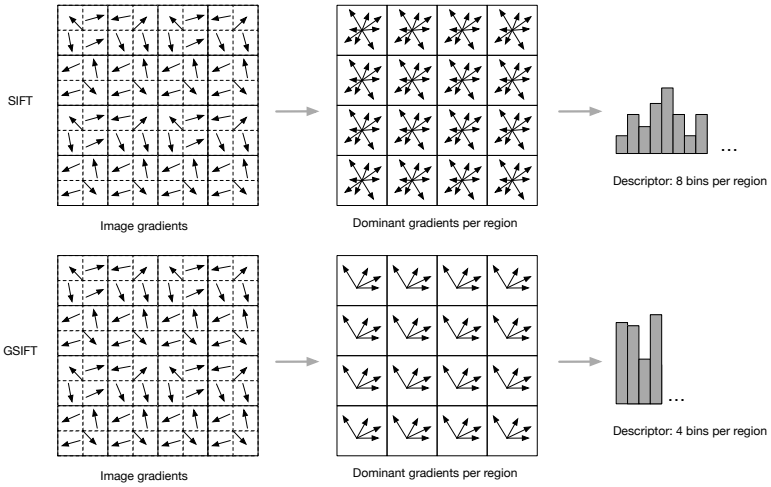


Figure 2.5: The SIFT and GSIFT feature descriptors.

putation of the contours, the process to build the histogram is the same as SIFT. Years later, this algorithm was extended by [41]. The authors decided to combine the shape information from EHD with frequency information using Log-Gabor coefficients (24 in total) at the center of each feature point, improving the performance of the previously described descriptor.

In addition to the works presented above, methods not based on feature descriptors have been also proposed to find correspondences between images of two different spectral bands. For example, [34] propose a non-rigid method to register a visible and an infrared image of a face. The proposed method consists in using edge maps from the images to represent the faces as an initial step, and later register, in a coarse-to-fine way, the edges using a technique based on the Gaussian field criterion. Another example is the work from Shen et al. [48], where the authors created a dense matching strategy based on variational approaches to match multi-spectral and multi-modal images.

It is important to note that although several methods have been suggested to describe multi-spectral images, the feature matching performance is still low in comparison with the performance of classical feature descriptors used to describe visible images. Thus, improve this situation is one of the main motivation of our work that will be considered in further Chapters.

2.2.2 CNN-Based description

For many years, carefully hand-designed feature descriptors, such as SIFT [33] and SURF [6], have been the key component of many and diverse vision tasks. However, in the last few years, such approaches have been started to be outperformed

by CNN-based solutions in multiple benchmarks (e.g., [4, 50, 59]). Broadly speaking, convolutional neural networks (CNN's) are a type of feedforward neural network that have become extremely popular since 2012, thanks to the results achieved by Alex Krizhevsky in the ImageNet competition of that year [28]. Krizhevsky et al. set a new record in the object classification challenge, reducing the top minimum error rate, up to the date, from 26% to 15%. As a consequence, from that day on, CNN's have been progressing at an incredible speed outperforming previous solutions in almost every area of active research.

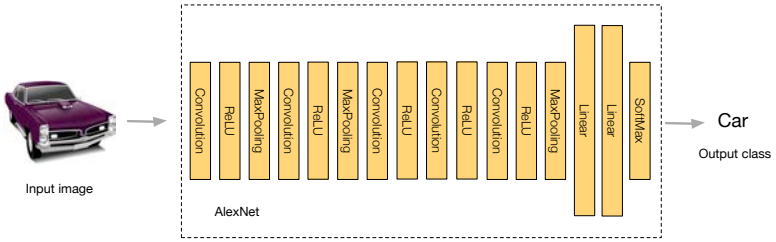


Figure 2.6: AlexNet network model.

One of the first works on CNN-based descriptors is the work of Fisher et al. in 2014 [16], named *Descriptor Matching with Convolutional Neural Networks: a Comparison to SIFT*. In their work, the authors analyzed the performance of trained convolutional filters, from the ImageNet challenge, as a replacement of hand-engineered feature descriptors. Essentially, they describe image patches as the output of the fourth convolutional layer of the AlexNet network model (see Figure 2.6), overlooking higher layers. Unexpectedly, such approach outperformed SIFT in the majority of the categories of the benchmark made by Mikolajczyk et al. [36]; which is considered the most popular test for this kind of evaluations. Moreover, it showed that filters specialized to discriminate between object classes could recognize subtle structures that are important to feature matching. The only drawback of this method was regarding the computational cost of computing the descriptor; SIFT was at least four times faster.

Following the success of the above method, custom network model specifically designed and trained to be used as feature descriptors started to appear. In this line, Zagoruyko et at. [59] designed and evaluated several different CNN-based network models to describe and match image patches. In contrast to [16], instead of comparing the resulting descriptors using an L_2 metric, they used a fully connected layer to learn a custom metric. Essentially, their training method rewards small distance difference between similar patches and huge difference between non-similar image patches. Results showed the validity of their different proposals. However, using a custom learned metric to compare patches is not efficient in many applications that need to compare thousands of features per second. Additionally, classical fast nearest neighbor strategies to improve the matching speed, such as KDTree, cannot be used, since the custom metric does not meet the triangle inequality property. A smart solution to this problem is presented in [23]. In their work, the authors

propose to separate the feature descriptor networks models in two parts, one for feature extraction and another for metric learning. Separating both functions leads to better performance since it is possible to compute all the features just once, and then using the metric network to match the features. Although this solution improved the performance of the previous solutions, still could not use strategies like KDTrees to improve the matching speed.

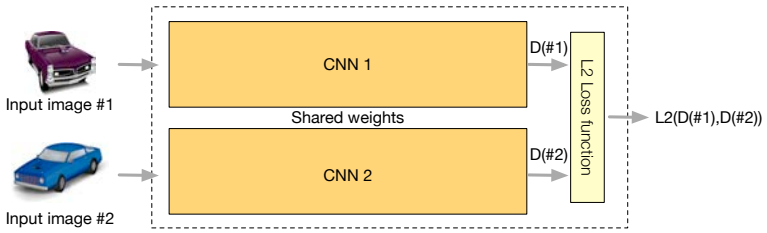


Figure 2.7: Siamese network architecture used by [50]

Since previous network models rely on custom metric networks, it was impossible to use them as a drop-in replacement of traditional feature descriptors. The work of Serra et al. [50] changed this circumstance. Serra et al. followed a similar approach to [59], but instead of using a custom metric network to measure the distance between two image patches, it directly minimizes the L_2 distance between them at training time. Their training architecture, depicted in Figure 2.7, consists of a siamese network with an L_2 loss function, where each CNN works as a feature descriptor and the loss function encourages small L_2 distances between similar patches. At testing time it is just necessary one of the two CNN to compute the descriptor of an image. Each descriptor calculated in this way can be used as a drop-in replacement of traditional feature descriptors, such as SIFT. Additionally, the authors make another significant contribution. During the training stage, they notice that after a few training epochs, most of the non-matching samples used to train the network were not giving new information, making it necessary to use mining strategies to improve the performance of the networks.

From [50], it was clear that an efficient use of the training samples leads to an increase in the overall performance of the trained feature descriptors. Inspired by the previous results, [4] proposes a triplet training network architecture to mine non-matching patches within each training sample. The triplet training network consists of three copies of the same CNN and a L_2 loss function that makes use of a triplet of input patches where two correspond to a matching example and one to a non-matching sample to the other two patches. Essentially, what the network does is to select at each training step one matching sample and the most difficult non-matching sample, from the two possible options. Results showed the validity of their proposal regarding feature matching performance and training speed. An efficient use of the matching samples improves the overall training speed in contrast to previous CNN-based techniques. It is important to notice that triplets have many uses in different contexts, e.g., [47] uses a triplet network architecture to recognize faces.

Our network proposal in Chapter 6 is strongly based on the triplet network proposed by [4], but adapted to be used in cross-spectral scenarios. We use quadruplet instead of a triplet, and we do not only mine non-matching samples but also correctly matched samples.

CNN-based descriptors have proved to be not only valuable to find feature correspondences between different images but also in additional vision tasks. For instance, [7] proposes a new method to track objects in a video sequence using patch similarity. Essentially, the authors train a network in a similar way to the above-described methods and later track objects matching candidate regions in following frames. Their strategy proved to have several advantages compared to the state-of-the-art. Firstly, they accomplished to achieve a tracking speed superior to 80 frames per second, something impossible to that date using CNN-based solutions. Secondly, their proposal did not need labeled videos sequences to train their network. Since the technique compares regions of interest between frames, static images can be used to train the network, simulating movement through rotation and translation over objects.

2.2.3 Cross-spectral applications

The number of cross-spectral applications, based on the usage of natural images from the visible and infrared spectral band is limited; with natural images refers to non-medical images and images not captured from satellites (remote sensing). Mostly due to the cost of cameras from spectral bands beyond the visual spectrum, the small number of public datasets available and the challenges related to acquiring such type of registered data, required to validate new algorithm proposals. Typically, applications fall into two categories: face recognition and video surveillance.

Face recognition in heterogeneous environments is one of the most attractive application of cross-spectral imagery. It consists of matching image faces from one spectral band to another. In [26], the author analyzes the advantages and limitations of matching images of faces taken from a visible camera to faces taken from a SWIR camera. The motivation behind their study is that information from the visible cameras can be complemented with information from the SWIR spectral band cameras in harsh environmental conditions, such as low-light and fog. Essentially, the proposed system consisted in describing faces using standard feature descriptors and then matching the faces using their vectorial representation. Their experiments demonstrated the suitability of such approach. However, they did not try custom cross-spectral descriptor, such as the one described in the previous section, that could have improved the feature matching performance. Other work on the same line but using different spectral bands is the one presented in [25]. In their work, the authors use learning techniques to recognize images of faces from a visible camera to faces from an NIR camera. Figure 2.8 shows image samples from a well known VIS-NIR faces database.

Other types of applications include the recovery of depth information from cross-spectral stereo rigs, such as the ones that can be found in high-end video surveillance

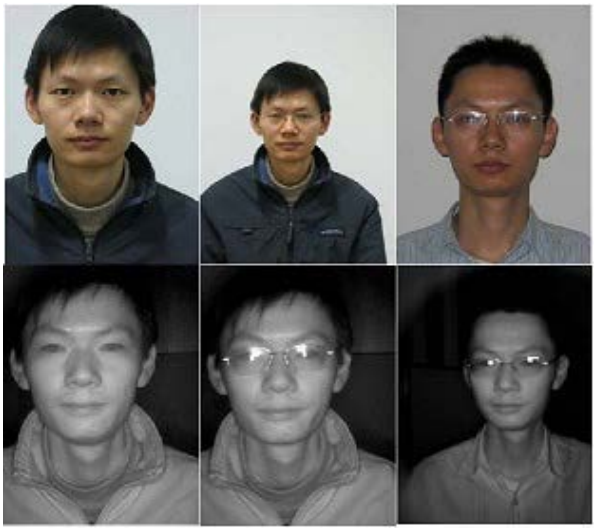


Figure 2.8: Image samples from the CASIA NIR-VIS 2.0 database. Top row, visible images, and bottom row, NIR images.

systems. Krotostky and Trivedi in [29] examine, develop and evaluate applications of cross-spectral stereo for pedestrian detection and tracking, recovering depth information from the cameras using mutual information as the matching cost function. Results showed the potential of such applications. However, also show that mutual information was not suitable as a matching cost function between images from the visible and LWIR spectral bands. Later, Pinggera et al. [43] focused on the same problem that Trivedi left open, but of using mutual information as the cost function, the authors used a modified version of the HOG descriptor, improving all previous results. It is important to notice that, up to the date, the computation of cross-spectral stereo disparity is still an open problem that has not been successfully tackled. However, we believe that in a few years this will change, thanks to contributions of new cross-spectral stereo datasets, such as [53].

In Chapter 7 we present a novel application of cross-spectral imagery. We use a visible camera and an LWIR camera mounted in a stereo-rig to localize the position of a car at every moment using visual odometry techniques.

Chapter 3

Benchmarks

The usage of several sets of images is an essential component of every research work in order to evaluate and compare new proposals. Unfortunately, there is a limited number of cross-spectral benchmarks available in the literature. Therefore, the acquisition of multiples sets of images from different spectral bands was an essential component of this work. With the new benchmarks, we were able to train, test and evaluate the proposals that we make in following chapters. Hence, the objective of this chapter is to describe the different benchmarks that have been created during this thesis; including from the acquisition process to the final resulting images.

3.1 Introduction

Ideally, every study published in a scientific article should be replicable by their peers; this is essential in any experimental science. However, many published works are not easy or even impossible to replicate due to lack of detailed explanations, data, and even bad practices. For that reason, the computer science community is actively encouraging the access to data and code to be able to reproduce results from scientific articles, and even more important, to make a fair and accurate comparison between the different proposed methods.

A major factor in the reproducibility of the results is the existence of open benchmarks. A benchmark can be described as the combination of data with a defined evaluation methodology that allows fair comparisons between different techniques proposed by scientists. Even more, it helps to reduce the funding differences that can exist between educational institutions, allowing researchers to experiment with data, that otherwise, would be impossible to obtain using their own resources.

In this chapter, we introduce different methodologies that exist to evaluate the performance of feature descriptors. As discussed in previous chapters, feature descriptors are compact representations of images, which are core in many computer vision

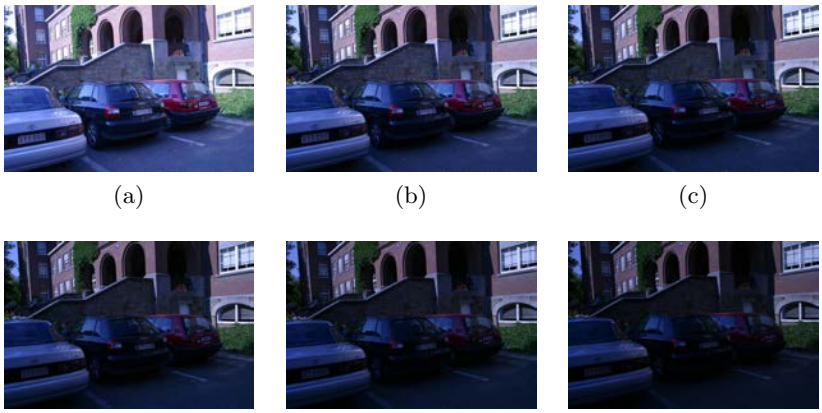


Figure 3.1: Leuven (light) sequence from the Oxford dataset.

application. Hence, the validation of the performance of each proposed method is essential to make a proper use of them later. Additionally, we introduce two benchmarks generated in the development of this thesis in an analogous way to the benchmarks described before. Finally, we also introduce a VIS-LWIR cross-spectral visual odometry benchmark, that we used to test a real application of cross-spectral feature descriptors. Broadly speaking, visual odometry consists in determining the pose, positions and orientation, of an object using images from cameras relative to their initial reference system.

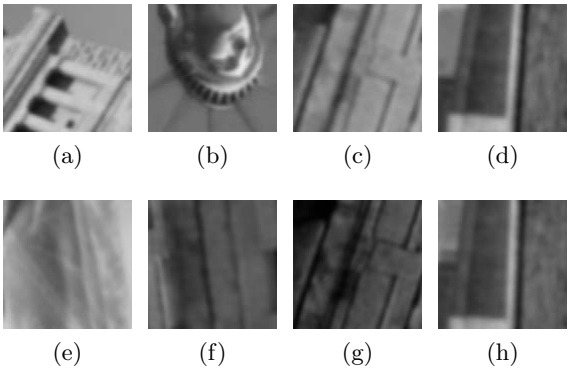


Figure 3.2: Image pair samples from the phototour image patch dataset. Two non-matching samples: (a)-(e) and (b)-(f) and two matching samples: (c)-(g) and (d)-(h).

3.2 Feature descriptor benchmarks

Different benchmarks have been proposed to evaluate the performance of feature descriptors. Coarsely speaking, they can be classified into two categories: *i*) benchmarks where feature descriptor performance is obtained from a local feature matching scheme—pairwise image registration; and *ii*) benchmarks where feature descriptor performance corresponds to an evaluation of similarity between image pairs.

The evaluation of a feature descriptor in a feature matching scheme occurs as follows. First, a feature detector is applied to each image, extracting multiple interest points as a result. The extracted feature points are defined by their position and size, the latter described by an ellispe. Then, a feature descriptor is applied to each feature extracted in the previous step, obtaining a compact, float or binary, representation of each interest point. Following, resulting descriptors from both images are compared by a distance metric, usually L_2 distance for floating descriptors and Hamming distance for binary descriptors. Finally, the matched descriptors are compared with ground truth data, to determine the overall performance of the technique.

The dataset of Mikolajczyk et al. [36] is an example of a feature matching based benchmark. It contains forty-eight images, divided into eight sequences of six images each. Each sequence is intended to cover common challenging matching scenarios, two sequences with blurred images, two sequences with changes in the viewpoint, two sequences with scale and rotation differences, one with JPG compression, and one with different lighting conditions (see Figure 3.1).

On the contrary to the previous approach, benchmarks based on a distance evaluation are used to estimate the degree of similarity between two images.In this context, two images are similar if the distance in the feature descriptors space is small, and different if the distance is big; the opposite case is also valid, big distances could indicate similarity and small distances could indicate dissimilarity. Data samples can be of two types: matching and non-matching. Matching samples are pairs of similar images, showing the same scenarios from a different perspective, or scale. Non-matching samples are randomly selected image pairs that are not visually related. The performance is measured by generating a ROC curve, from the similarity measures of all the test samples in the dataset. In other words, the performance metric measures how well the descriptor discriminates similar patches from the dissimilar ones.

The dataset from Winder and Brown [57] is the most popular in this category of evaluation. The dataset contains more than 1 million of different image patches of 64x64 divided into three categories: liberty, Notre-Dame, Yosemite. Each category name indicates the provenance of the image samples. Figure 3.2 shows sample images from the dataset. This dataset will be used in Chapter 6 to validate our proposal, which not only works well on cross-spectral scenarios but also in visible to visible cases.

In the rest of the section we will describe the benchmarks acquired and generated during this thesis.

3.2.1 VIS-LWIR feature matching benchmark

This section describes the dataset of VIS-LWIR cross-spectral image pairs that we have collected to train and evaluate cross-spectral feature descriptors, in a similar way to [36]. The process consisted in developing a cross-spectral stereo-rig and taking outdoor images. These two steps are described next, indicating the used hardware, the challenges of calibrating cross-spectral images and the resulting images.

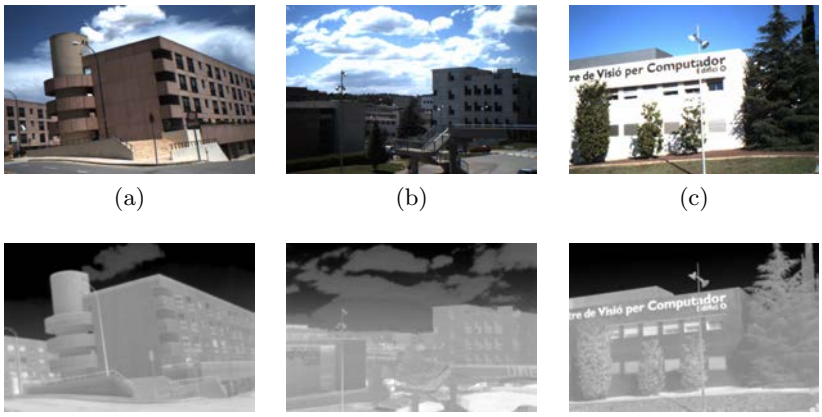
Figure 3.3: Multi-spectral stereo-rig.

Hardware configuration

The platform used to collect the images consisted of two cameras, one that captured the visible band and one that captured the thermal infrared band. The camera used to obtain images from the visible spectrum (Basler ACE acA645-100gc) captures up to 100 *fps* with a resolution of 658×492 pixels. The thermal camera (Xenics Gobi-640-GigE) captures up to 50 *fps* with a resolution of 640×480. The focal lengths of the cameras were set so that pixels in both images contain almost the same amount of information of the observed scene. An image-rig on top of a tripod was used to mount both cameras. Additionally, an external trigger was used to synchronize the acquisition of images from both cameras.

Data collection

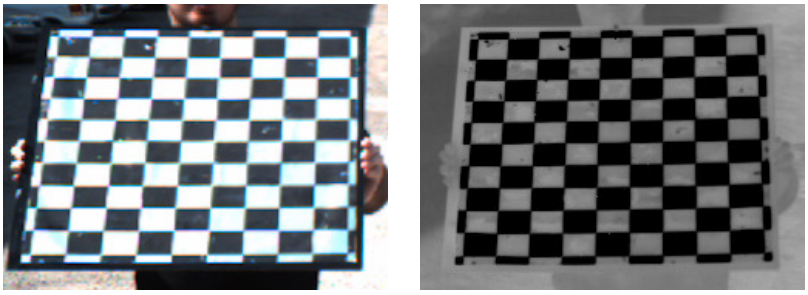
The obtained dataset consists of 44 VIS/LWIR images pairs that were captured using the multi-spectral image-rig described above (Fig 3.3 shows an illustration of the acquisition hardware). All image pairs were captured in outdoor locations around the Autonomous University of Barcelona's campus; the obtained scenes mostly contain buildings and vegetation. Figure 3.4 shows image pairs samples from the acquired dataset. In the dataset each image pair is registered, so both images are in a common coordinate frame. The description of the method used to register images from both spectral bands is detailed next.



(a) (b) (c)

(d) (e) (f)

Figure 3.4: VIS-LWIR image pairs; top images are from the visible spectrum and bottom images from the LWIR spectrum.



(a) (b)

Figure 3.5: (a) Calibration pattern image from the visible spectrum. (b) calibration pattern image taken with the thermal camera.

Calibration

To calibrate each camera we have constructed a chessboard calibration pattern visible to both cameras, visible and thermal. Typical calibration patterns are not visible on thermal cameras. Although a difference in color may exist between the squares of a chessboard, not necessarily it also exist in temperature. To overcome this problem, we printed a chessboard calibration pattern on top of a reflective metal plate that in front of the sun light it becomes visible in both spectral bands. Figure 3.5 shows two images of our calibration pattern, one from the visible spectrum camera, and one from the thermal camera

Once we had a calibration pattern visible to both cameras, we proceeded to cal-

ibrate the cameras using Bouguet's calibration toolbox [9]. As a result, the intrinsic and extrinsic parameters from both cameras were found. Later these parameters were used to rectify all the images.

Finally, images were aligned manually, selecting corresponding points in both images to compute the homography that relates both images. Although, in most cases, it was not necessary, since images were taken far away from the image-rig and after the rectification, most of the images were already aligned, i.e, the computed homography matrix was close to the identity matrix.

3.2.2 VIS-NIR image patch benchmark

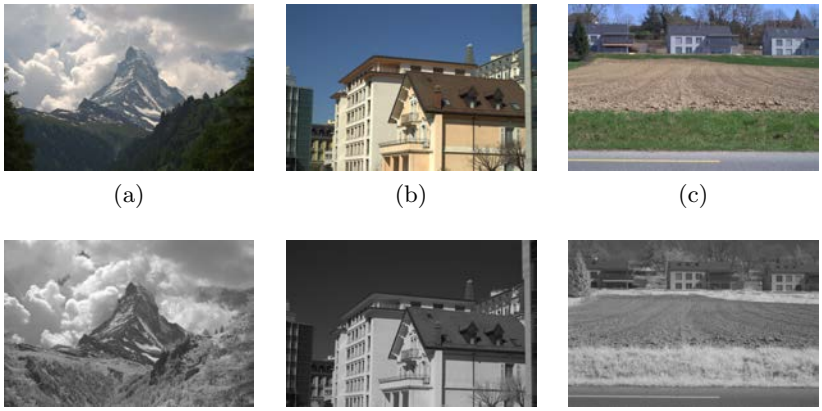


Figure 3.6: Images samples from the Oxford dataset. The six images correspond to the different illumination category.

Deep learning solutions require a huge quantity of data to train their network models. Therefore, to train, test and evaluated the proposed method in Chapter 6, we built a cross-spectral image patch dataset using the public VIS-NIR set of images from [15]. The later set consists of 477 VIS-NIR image pairs divided into nine different scene categories: country, field, forest, indoor, mountain, old building, street, urban, and water. Images are in TIFF format at 1024×768 resolution. Image pairs are registered.

We generated the set of image patches as follows:

1. Interest points were detected in the original visible images using SIFT.
2. Image patches of size 64×64 were extracted centered in the interest points detected in the previous step.
3. Half of the patches were selected to be positive samples (matching patches), and half to be false samples (non-matching patches).

4. For the matching patches, we looked for the same interest points in the original NIR image, and subsequently extracted a patch of 64×64 in that location. Since images are registered, the interest point's location in images from both spectral bands is the same.

5. For the non-matching patches, we selected a random patch of size 64×64, far from the original patch in the NIR image.

As a result, we obtained more than one million of VIS-NIR image pairs. Figure 3.7 shows image patch samples from the dataset and Table 3.1 the detailed number of image patches per category.

Category	# patches
country	277504
field	240896
forest	376832
indoor	60672
mountain	151296
oldbuilding	101376
street	164608
urban	147712
water	143104

Table 3.1: Number of cross-spectral patches per category.

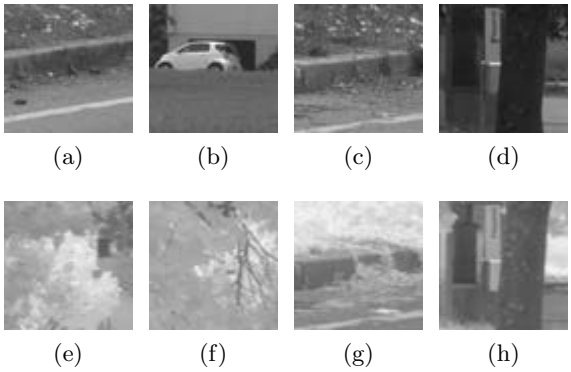


Figure 3.7: Samples of pairs of images from the VIS-NIR image patch dataset. First row corresponds to grayscale images while second row shows the corresponding NIR images—non-matching samples: (a)-(e) and (b)-(f); matching samples: (c)-(g) and (d)-(h).

3.3 Visual odometry Benchmarks

As stated before, visual odometry consists in estimating the pose of an agent over time, using one or multiples cameras. It is an essential component for motion tracking, obstacle detection and avoidance; mobile robots and autonomous cars, navigation systems, just to mention a few. Although several methods have been proposed in the literature to estimate the change in position and orientation over time (odometry) without the use of cameras, such as GPS, laser, and INS systems, visual odometry system excels in obtaining a high-accuracy localization at a low value; compared to solutions that uses expensive hardware such as LIDARs. Additionally, visually based methods work well on multiples scenarios like indoor and outdoor areas, even underwater, which is not the case with some of the technologies mentioned above.

A visual odometry benchmark consists of a collection of videos, recorded from a moving agent perspective, where the actual pose of the agent, at every instant, is known and later presented in the form of ground-truth data. For instance, Sturm et at. [51] proposed a RGB-D benchmark for the evaluation of visual odometry algorithms on mobile robots using color and depth information. The agent, a mobile robot Pioneer P3DX with a Microsoft Kinect camera on top, captured color and depth information at each step, from the initial position to the final one. Video sequences were recorded in a controlled indoor environment, under different conditions, such as pedestrians walking close to the robot. Ground-truth pose estimation was obtained using eight high-speed tracking cameras from a commercial motion tracking system, delivering a high-quality pose data at each step. Another popular benchmark is the one proposed by Geiger et al. [21]; the dataset, named, KITTI benchmark, consists of twenty-two stereo sequences, acquired using as a platform a Volkswagen Passat B6 with multiples sensors on board. On the contrary to the previously described benchmark, KITTI is intended to evaluate algorithms that are useful for an autonomous driving assistant system or, in general, autonomous cars.

The relative pose error (RPE) is used to measure the performance of visual odometry systems. RPE measures the accuracy of the computed poses over a fixed period of frames. For example, assume that $P_1, ..., P_n$ are the poses estimated by a proposed algorithm, and $Q_1, ..., Q_n$ the real poses provided by the benchmark, then the RPE is defined as follows

$$RMSE(E_{1:n}) = \frac{1}{n} \sum_{\delta=1}^{n} RMSE(E_{1:n}, \Delta)$$

where $E_{1:n}$ indicates the relative pose error at each time step i and Δ the selected time-interval.

In this thesis, we acquired a VIS-LWIR benchmark to evaluate cross-spectral stereo visual odometry methods. The benchmark is described next.

Figure 3.8: Electric car used to capture the VIS-LWIR multi-spectral odometry dataset. Cameras are mounted on the roof.

Figure 3.9: Images samples from the visual odometry dataset

3.3.1 VIS-LWIR benchmark

With the intention of evaluating real applications of cross-spectral feature descriptors, we proposed a VIS-LWIR cross-spectral visual odometry benchmark.

Hardware configuration

The cameras and stereo-rig are the same than the one described in Section 3.2.1, with the difference that this time the image-rig was mounted on the roof of an electric car, instead of a tripod. The separation between the cameras was about 12 cm. Additionally, since in visual odometry we care about the position of the moving platform, we added a low-cost GPS tracking system to obtain the position of the car at every movement. The GPS updated its coordinates almost two times per second, which can be considered as a limitation of this dataset since the update speed of the GPS is much slower than the camera's framerate. Figure 3.8 shows an image of the electric car with the cameras mounted on the roof.

Data collection

Five sequences of videos were captured, split up into semi urban and rural scenarios detailed in Table 3.2. The former are richer regarding visual characteristics than the latter. However, at the same time, they present more probabilities of containing nonstationary objects (i.e., vehicles, pedestrians...). All these sequences represent real traffic conditions with strong illumination variations and lack of texture. The texture issue applies more to thermal images and can be explained by the fact that LWIR pixel brightness depends on heat variations. Most of the lower part of images is composed by ground surface, where heat does not vary a lot. This means that this part of the image would be textureless and therefore not useful for visual odometry methods based on local features matching. Figure 3.10 shows the trajectories of each sequence on the map.

Video #	Scenario Type	Traveled Distance (m)	Average Speed
Vid00	Urban	240	17 Km/h
Vid01	Urban	470	23 Km/h
Vid02	Urban	450	20 Km/h
Vid03	Rural	350	30 Km/h
Vid04	Rural	260	25 Km/h
Total		1770	

Table 3.2: Multispectral video sequences

(a) Vid00 (b) Vid01

(c) Vid02 (d) Vid03

(e) Vid04

Figure 3.10: Video sequence trajectories. Yellow arrow indicates the starting point of the car.

Chapter 4

Log-Gabor based cross-spectral description

In this chapter, we propose a new hand-made cross-spectral feature descriptor, which is suitable to find correspondences between images from different spectral bands or modalities. The descriptor, referred to as Log-Gabor Histogram Descriptor (LGHD), describes the neighborhood of feature points combining frequency and spatial information using multi-scale and multi-oriented Log-Gabor filters.

4.1 Introduction

As discussed in Chapter 2, finding correspondences between images from different spectral bands or modalities is a challenging task. For that reason, in this chapter we propose a novel hand-made feature descriptor suitable to the task of matching features points between images with nonlinear intensity variations. This includes image pairs with significant illuminations changes, cross-modal image pairs, and cross-spectral image pairs. The proposed method describes the neighborhood of feature points combining frequency and spatial information using multi-scale and multi-oriented Log-Gabor filters.

The rest of the chapter is organized as follows. The proposed descriptor is introduced in Section 4.2. The evaluation methodology together with the evaluation results are presented in Section 4.3. Finally, conclusions are given in Section 4.4.

4.2 Proposed approach

The nonlinear intensity variations between a pair of images can be the result of different configuration setups, which can affect each image differently. However, de-

spite these intensity differences, the global appearance and the shape of the objects contained in the scene tends to remain constant. This fact makes us think that a descriptor based on the distribution of high-frequency components would be robust to different nonlinear intensity variations, which is the idea behind the proposed approach.

The current work is motivated by the progress made by the local EHD descriptor presented by Aguilera et al. [1], in contrast to previous approaches. The EHD descriptor describes the spatial edge distribution around a point computing an orientation histogram of eighty bins. For each interest point, a region of $S \times S$ is defined and further divided into sixteen smaller sub regions (4×4). Within each subregion an orientation histogram of five bins is computed using the strongest pixel value for one of five different oriented Sobel filters (horizontal, vertical, 35 degrees, 135 degrees and non-oriented).

The proposed Log-Gabor Histogram Descriptor (LGHD) describes local patches in a similar way to EHD, but instead of using multi-oriented Sobel descriptor it uses multi-oriented and multi-scale Log-Gabor filters.

4.2.1 Log-Gabor filters

Broadly speaking, a Log-Gabor filter is an image processing tool that can be used to obtain localized frequency information in an image. In other words, Log-Gabor filters can decompose an image in terms of frequency responses at different scales and orientations. Log-Gabor filters are the keystone of several computer vision algorithms (e.g., [27, 41]).

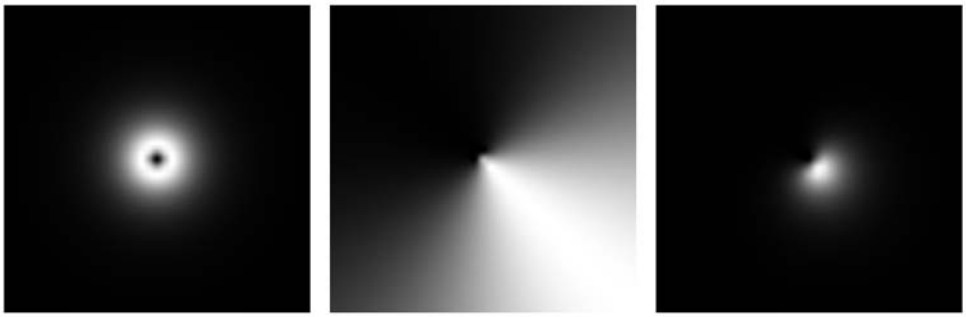
Log-Gabor filters have several useful properties. Filters can be constructed with any arbitrary bandwidth and, by definition, they do not have a DC component. Thus, the response of the filter do not depend in the mean value of the signal; in this case an image. Therefore, Log-Gabor filters are more robust to different lighting conditions than other frequency analysis tools like Gabor filters.

Formally, a 2D Log-Gabor filter is described as Equation 4.1, where f_0 is the central frequency, σ_f is the frequency width, θ_0 is the center orientation and σ_θ the width of the orientation component. Consequently, each filter is composed by two elements; one based on a frequency and another based on an orientation. Equation (4.2) shows how to compute the frequency bandwidth and Equation (4.3) the angular bandwidth. See Figure 4.1 for a visual example of a Log-Gabor descriptor.

$$G(f,\theta) = exp\left(\frac{-(log(\frac{f}{f_0}))^2}{2(log(\frac{\sigma_f}{f_0}))^2}\right) exp\left(\frac{-(\theta-\theta_0)^2}{2\sigma_\theta^2}\right) \tag{4.1}$$

$$B = 2\sqrt{\frac{2}{log(2)}}(||log(\frac{\sigma_f}{f_0})||) \tag{4.2}$$

$$B_\theta = 2\sigma_\theta \sqrt{2log2} \tag{4.3}$$



(a) Frequency (b) Orientation (c) Combined filter

Figure 4.1: A Log-Gabor filter is the combination of a frequency component with an orientation component.

It is important to notice that there is not a unique or ideal arrange of Log-Gabor filters. Thus, the design of a filter arrange is application oriented and often more art than science.

4.2.2 Histogram descriptor

The proposed histogram descriptor can be computed following the next steps:

1. Create a filter bank and convolve each Log-Gabor filter with the input image patch. In this thesis, we use twenty-four filters at six orientations between $[0, \pi]$ in four different scales. As result, twenty-four image patches are generated, where each one represents the frequency and orientation response to one of the filters from the bank.
2. Divide the image patch into sixteen sub regions (4×4).
3. Compute the dominant orientation at each pixel using the magnitude of the filters response at the first scale in the first sub region; repeat for the other scales and sub regions.
4. Build a histogram of oriented Log-Gabor filters in each sub region using five bins: one per orientation. Repeat this process to all the other regions and scales.
5. Concatenate all the histograms from the previous step. The resulting histogram is 384 bin long (96×4).

Figure 4.2 shows a visual illustration of the proposed approach.

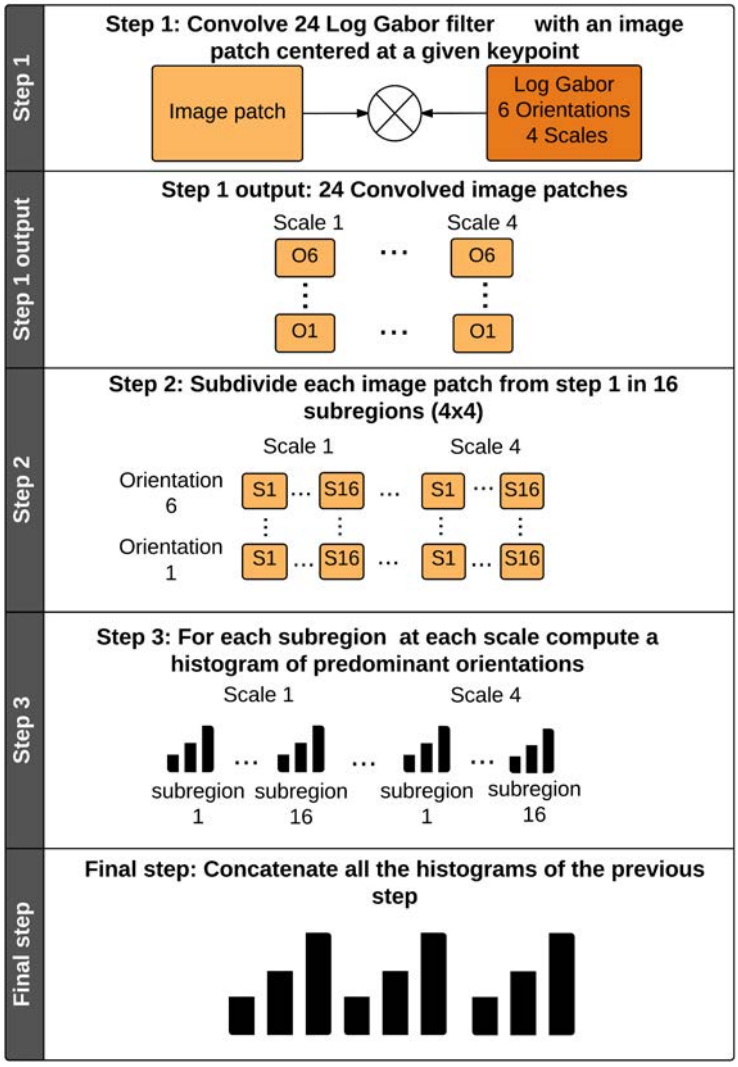


Figure 4.2: Illustration of LGHD steps.

4.3 Experimental evaluation

The proposed approach has been evaluated using four different set of image pairs: 58 RGB/NIR pairs taken from the urban sequence from the benchmark described in Chapter 3; 44 RGB/LWIR outdoor pairs from our VIS/LWIR benchmark; 120 FLASH/NO-FLASH pairs of images from [49] and 4 RGB/DEPTH pairs from [48]. Figure 4.3 shows sample image pairs of each benchmark.

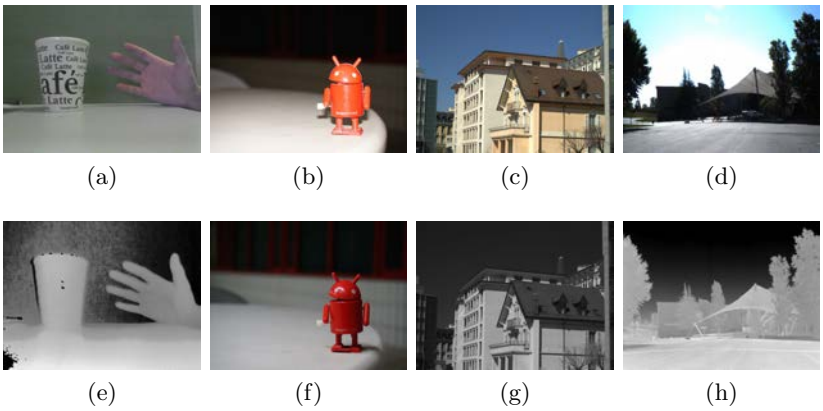


Figure 4.3: Examples of pairs of images from the four data sets evaluated in the current work. **This figure is best viewed in color**.(a)-(e) RGB/DEPTH image pair, (b)-(f) FLASH/NO-FLASH image pair, (c)-(g) RGB/NIR image pairs, and (d)-(h) VIS/LWIR image pairs.

In addition to the evaluation mentioned above, the proposed approach has been compared with four state-of-art descriptors: *1)* the EHD descriptor that was originally proposed for the RGB/LWIR case [1]; *2)* the gradient invariant version of SIFT (GISIFT) [14]; *3)* the PCEHD descriptor [41]; and *4)* the SIFT descriptor [33] that is used as a reference of classical descriptors.

In order to evaluate the performance of feature descriptors, avoiding bias due to feature detector performance, we follow a similar approach to [10]. We detect features just in one image using the FAST detector [46], and then we project them into the corresponding pair using the homography information. This process is done for 3 sets of the image pair; for the remaining one (RGB/DEPTH) we use 100 points manually selected (provided by [48]), since the images cannot be represented by a unique homography.

The performance of the different descriptors is evaluated using the resulting matching precision:

$$Precision = \frac{C}{T}, \tag{4.4}$$

where C is the number of correct matches and T is the total number of correspondences.

In our experiments we convolve the different images with Log-Gabor banks using the Matlab implementation of [27]. We set *nscale=4*, *norient=6*, *minWaveLength=3*, *mult=1.6* and *sigmaOnf=0.75*. Additionally, Table 4.1 shows the different patch sizes used to evaluate the EHD, PCEHD and LGHD descriptors (these sizes were empirically obtained in order to have a fair evaluation). The matches are found by

using Euclidean distance (SSD).

Descriptor	RGB/DEPTH	Other cases
EHD	32×32	80×80
PCEHD	32×32	80×80
LGHD (Ours)	32×32	80×80

Table 4.1: Patch sizes used to evaluate the EHD, PCEHD and LGHD descriptors.

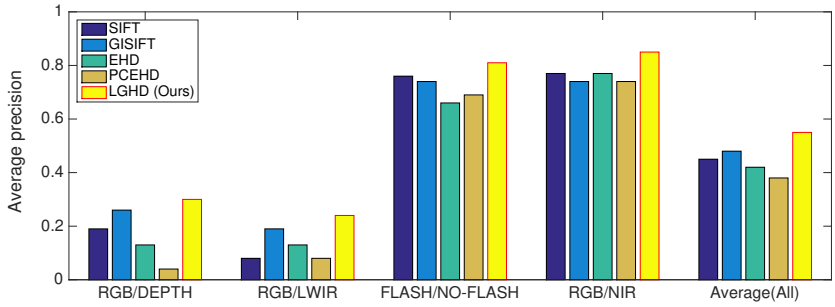


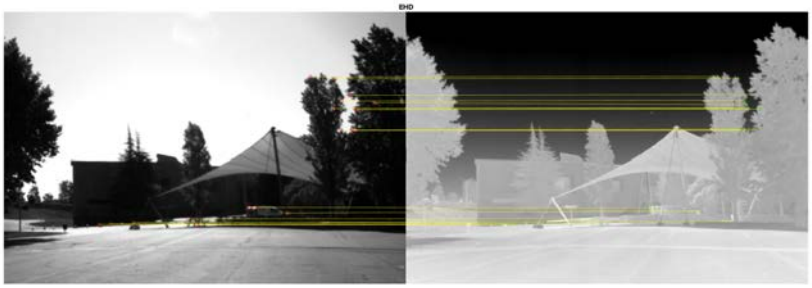
Figure 4.4: LGHD results in four different datasets.

Results are shown in Figure 4.4, where each bar of the graph indicates the average matching precision for the corresponding descriptor computed over the whole data set. The proposed approach, LGHD, obtained the best performance in every category when compared with all the other descriptors evaluated in the current work. Regarding computational times, the proposed LGHD descriptor has a similar performance to other approaches with respect to the feature description estimation, but its matching cost is the most expensive one due to the size of the description vector (384 elements).

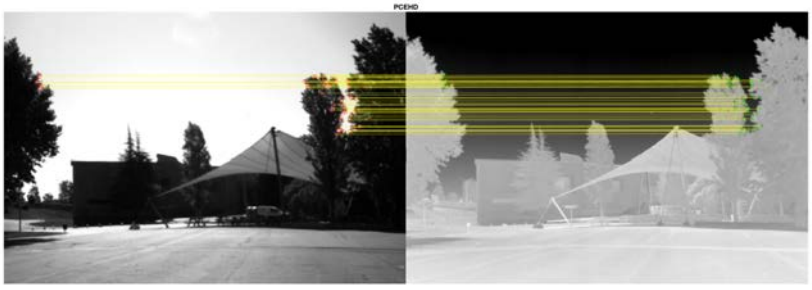
The matching precision for the FLASH/NO-FLASH and the RGB/NIR cases was considerably higher than in the other two scenarios. This fact is mainly due to the spectral band closeness of the image pairs: *i)* the NIR spectrum is the closest infrared band to the visible spectrum; *ii)* while in the FLASH/NO-FLASH dataset, the pairs of images correspond both to the same spectral band (the visible one). On the other hand, lower precision rates were obtained for the RGB/DEPH and RGB/LWIR cases. The LWIR band is the most distant infrared band from the visible spectrum. Image pairs mostly share shape information, while most of the texture information is missed. The depth case is even worse since all texture information is missed; in this case just a limited number of visual similarities between visible and depth images is kept.

Visual matching examples, of just one image pair (RGB/LWIR), are presented in Figure 4.5 for the three histogram based descriptors. In this pair of images, it

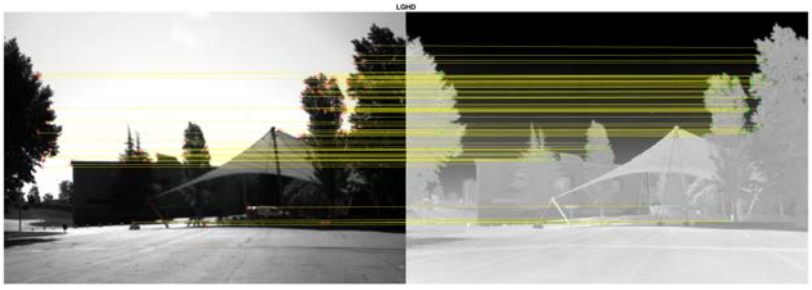
can be appreciated the difficulty of finding local similarities due to the large intensity variation between the images.



(a) EHD



(b) PCEHD



(c) LGHD

Figure 4.5: Resulting output after matching a visible image with an LWIR image.

4.4 Conclusions

In this chapter, a novel feature descriptor has been presented. It can be used to the task of matching features between images with non-linear intensity variations such as multi-spectral and multi-modal images. Results show that the proposed algorithm outperforms state-of-art algorithms in the four data sets considered in the evaluation. The RGB/LWIR and the RGB/Depth were the most challenging cases. Results show that in both cases matching descriptors generate an elevated number of mismatches. These mismatches can be reduced using a robust formulation such as RANSAC.

Chapter 5

CNN-based cross-spectral description

Although the performance of CNN-based descriptors in monocular matching scenarios is at the state-of-the-art, the performance and suitability of such network models are still unknown in cross-spectral scenarios. For that reason, in this chapter, we explore the usage of four different architectures proposed in the literature to describe image patches. Specifically, we train and evaluate each network architecture using images from the visible and near-infrared spectral bands, and then test against the two visible-infrared benchmarks described in Chapter 3.

5.1 Introduction

Recently, several new CNN-based descriptors learned from data have been proposed in the literature (e.g., [4, 50, 59]), showing improvements over traditional hand-made feature descriptors in different benchmarks regarding discriminative power. However, the performance in cross-spectral scenarios has been not yet tested. Thus, the objective of this chapter is to explore the suitability of CNN-based descriptors to measure the similarity between images patches from two different spectral bands; specifically in VIS-NIR and VIS-LWIR scenarios. Additionally, the following question is addressed: can CNN-based descriptors generalize well to spectral configurations different from the trained one?

The rest of the chapter is organized as follows. In Section 5.2 we describe each network model evaluated in this work, along with the train loss function to minimize. In Section 5.3 we describe the training procedure that was followed to train each network using a VIS-NIR dataset and the hardware environment used. The resulting evaluations are described in Section 5.4, and finally, conclusions are presented in Section 5.4.

5.2 Network architectures

Four CNN-based model architectures are considered in this study (i.e., 2-channel (2ch), siamese (siam), pseudo-siamese (psiam) and triplet models). The first three take as input two image patches at training and testing time, and the triplet model takes three input patches at training and two at testing time. We selected this four models mainly for two reasons

- Trained models are publicly available, so it is possible to compare their monocular trained model with our cross-spectral trained ones. This with the objective of answer the following two questions: *i*) can CNN-based descriptors trained in the visible domain be used in cross-spectral scenarios without modifications?—hand-made descriptors cannot. and *ii*) which is the difference between trained networks with visible monocular data compared to trained networks using multi-spectral data?.
- Secondly, these network architectures can be easily adapted for the cross-spectral case, just setting each network input to a particular spectral band. It is important to keep the setting at testing time

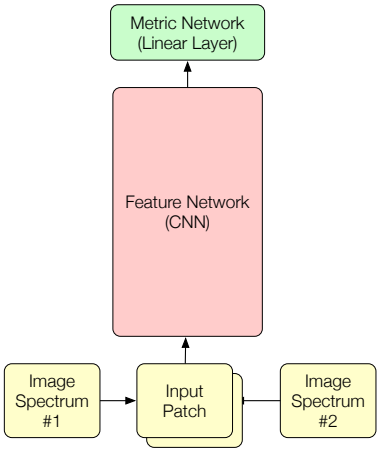


Figure 5.1: 2-ch network architecture

5.2.1 2-channel network

The 2-channel network model is depicted in Figure 5.1. It takes as input an image of two channels, where each channel corresponds to one of the two patches to be compared; one spectral band per channel. The network is composed of a series of convolution and normalization layers and a decision layer at the top that is trained to evaluate the similarity between the input patches.

The 2-ch network model architecture combines information of both spectral bands at the beginning of the feed forward process, processing in the next layers the combined information obtained in the previous steps. Processing the data jointly from the first layer has proven to be the best solution regarding feature matching performance in the visible monocular case. However, it has its drawbacks. Is one of the slowest solutions when matching local features, since it is not possible to reuse computed output layers, i.e., each patch pair is unique.

A margin criterion is used to train this network model. The margin criterion optimizes the two-class classification hinge-based loss term described by the following equation:

$$\min_{w} \frac{\lambda}{2}||w||_2 + \sum_{i=1}^{N} max(0, 1 - y_i o_i^{net}), \tag{5.1}$$

where w corresponds to the network weights, o_i^{net} is the network output for the i-th training sample, λ is the weight decay term and y is i-th training label. y can take two values, +1 if the i-th training sample is a correct match or -1 if it is a wrong one. In other words, the network is trained in such a way that we expect large positive values at the output of the network when both patches are the same and small values when the patches are different.

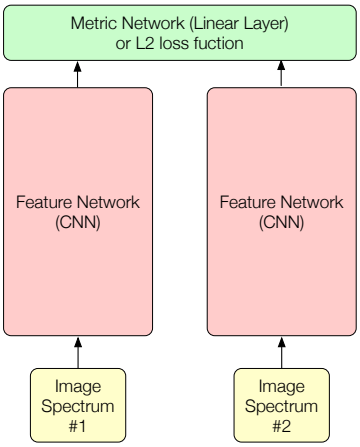


Figure 5.2: Siamese and pseudo-siamese network architecture.

5.2.2 Siamese network

In essence, a siamese network is quite similar to traditional feature matching approaches, i.e., the network firstly computes feature descriptors for each image patch

and then evaluates the similarity between the descriptions using some distance measure. The network consists of two CNN feature networks, which basically imitates a feature descriptor, with shared parameters that process each patch independently and a final decision network that acts as a distance metric (see Figure 5.2). Each feature network is composed of a series of convolutions, ReLU and max-pooling layers, while the metric network is composed of dense layers. It is important to notice that the final layer of the siamese network could be omitted at testing time, and replaced by a $p-$norm measure.

Siamese networks are slower than 2ch network at training but can be faster at prediction. This is mainly because, once trained, it is possible to divide the network into two different stages and separately compute the feature description from the similarity measure, i.e., it is plausible to reuse the output layers from previously introduced patches.

Although different loss function could be used with this type of network, we used the one described in the 2ch-network section. We did the same for the pseudo-siamese network

5.2.3 Pseudo-siamese network

The pseudo-siamese network is essentially a siamese network but without shared parameters, i.e., each feature network is different from the other. This is important since the pseudo-siamese network can end up learning custom convolutional filters for each input spectrum, giving more flexibility to the network. The setting used in the current work is the same than the ones used in the siamese network.

5.2.4 Triplet network

A triplet model is an architecture used at a training stage, since at testing stage it is the same than a siamese network. Figure 5.3 shows the training architecture of a triplet network. The network has three inputs, where each input corresponds to a different image patch. Formally, the input is a tuple $T = \{w, x, y\}$, where w and x are two matching image patches and y is a non-matching image patch to w and x. Each one of these patches will feed one of the three CNN *towers* that the network has; CNN 1, CNN 2 and CNN 3. The three CNN *towers* of the network share the same parameters during the entire training stage. Finally, the output of each *tower* will be a descriptor D of configurable size, that describes each input patch.

The loss function is described as follows:

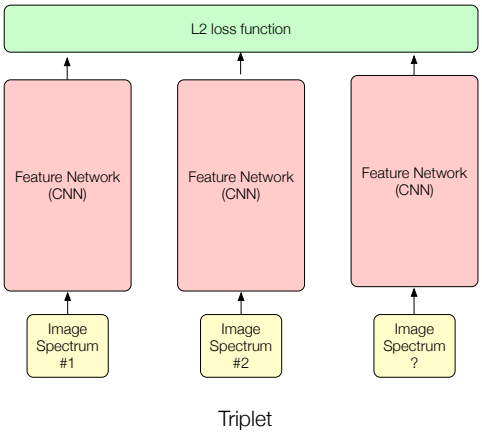


Figure 5.3: CNN architectures to describe or match similarities

$$P_m = \frac{e^{\Delta^+}}{e^{min(\Delta_1^-, \Delta_2^-)} + e^{\Delta^+}} \tag{5.2}$$

$$P_{nm} = \frac{e^{min(\Delta_1^-, \Delta_2^-)}}{e^{min(\Delta_1^-, \Delta_2^-)} + e^{\Delta^+}} \tag{5.3}$$

$$Loss(T_i) = P_m^2 + (P_{nm} - 1)^2 \tag{5.4}$$

where, Δ^+ corresponds to the L_2 distance between the descriptors of the matching pair w and x, $||D(w), D(x)||_2$; Δ_1^- corresponds to the L_2 distance between the descriptors of the non-matching pair w and y, $||D(w), D(y)||_2$; and Δ_2^- corresponds to the L_2 distance between the descriptors of the second non-matching pair x and y, $||D(x), D(y)||_2$.

In essence, the objective of the loss function is to penalize small L_2 distances between non-matching pairs, and large L_2 distances between matching pairs. Ideally, we want P_m to be equal to zero and P_{nm} to be equal to one, i.e, $\Delta^+ << min(\Delta_1^-, \Delta_2^-)$. Computing the minimum L_2 distances between the non-matching pairs is a type of mining strategy, where the network always performs backpropagation using the hardest non-matching sample of each triplet T, i.e., the non-matching sample with the smallest L_2 distance. The mining strategy is used to avoid the problems described in [50]. Finally, the mean square error is used to penalize values of P_m different of zero and values of P_{nm} different of one.

One key difference between monocular and cross-spectral image pairs is that for each cross-spectral matching pair we have two non-matching possible image patches; one for each spectrum. So the question is, which image patch we use as y?. We propose three simple and naive solutions: *i*) y is an RGB non-matching image, *ii*) y is an NIR non-matching image and *iii*) y is randomly chosen between RGB and NIR

images. In Section 5.4 we evaluate each alternative.

5.3 Training

All the networks described in the previous section were trained in a supervised way. To that end, we used the VIS-NIR cross-spectral image patch dataset described in Chapter 3.

The 2-ch, siamese and pseudo-siamese models were trained using Stochastic Gradient Descent (SGD) with a learning rate of 0.05, L2 weight decay (λ) of 0.0005, a momentum of 0.9 and batches of 256 samples. As recommended in [8], the training data was shuffled at the beginning of each epoch, and each input patch is normalized by its intensity mean. All the patches from the *country* category were used to train the networks, where 80% of the data were used as training data and 20% of the data as validation. Additionally, we augmented the training data flipping the cross-spectral image pairs horizontally, vertically and rotating both images in 90 degrees—to increase the training data and prevent overfitting.

The triplet network, also known as PN-Net [4], was train almost in the same way to the previously described networks, but with the following training parameter differences: learning rate of 1.1, weight-decay of 0.00001, batch size of 128, momentum of 0.9 and a learning rate decay of 0.000001.

The 2ch network parameters used in the current work are listed in Table 5.1. The siamese and pseudo-siamese feature networks have the same configuration than the 2ch network, just changing the metric network for the one described in Table 5.2. The PN-Net layer description is shown in Table 5.3.

All the code was implemented in Lua using the scientific computing framework Torch [12]. The hardware consisted of a 3.0 GHz Core I7 PC with a NVIDIA K40 GPU.

5.4 Experimental evaluation

Trained networks were tested with the two cross-spectral benchmarks presented in Chapter 3. In all the experiments presented below the networks are referred to as 2ch-country (2-channel network model trained on the country sequence), siam-country (siamese network model trained on the country sequence), psiam-country (pseudo-siamese network model trained on the country sequence), PN-Net RGB (PN-Net network trained on the country sequence using as third image a visible image), PN-Net NIR (PN-Net network trained on the country sequence using as third image a NIR image), and PN-Net RANDOM (PN-Net network trained on the country sequence using as third image a random image, that could be VIS or NIR). The country sequence was selected for training for two main reasons: *i)* in a preliminary evaluation stage the country sequence was one of the most difficult sequence in the VIS-NIR

Layer	Type	Output Dim	Kernel	Stride
1	convolution	96	7x7	1
2	ReLU	96	-	-
3	max-pooling	96	2x2	1
4	convolution	192	5x5	1
5	ReLU	192	-	-
6	max-pooling	192	2x2	1
7	convolution	256	3x3	1
8	ReLU	256	-	-
9	Linear	1	-	-

Table 5.1: 2ch network parameters.

Layer	Type	Input Dim	Output Dim
9	Linear	512	512
10	Linear	512	1

Table 5.2: Siamese and pseudo-siamese metric network parameters.

Layer	Description	Kernel	Output Dim
1	Convolution	7x7	32x26x26
2	Tanh	-	32x26x26
3	MaxPooling	2x2	32x13x13
4	Convolution	6x6	64x8x8
5	Tanh	-	64x8x8
6	Linear	-	256

Table 5.3: PN-Net layer descriptions

dataset; and *ii)* it is also one of the sequences with more data available. Additionally, for comparative evaluations with the state-of-art on VIS and cross-spectral patch similarity, we present results obtained with six of the trained networks presented in [59], SIFT [33], GISIFT [14], EHD [1] and LGHD [2].

5.4.1 VIS-NIR image patches

We evaluate the performance of our networks using the false positive rate at 95% Recall (FPR95) on each category of the VIS-NIR scene dataset (as in [59]). To be fair, we do not include the country sequence, since it was the sequence used to train

our networks

The results of our tests are presented in Table 5.4. All our networks perform better than the ones trained just with images from the visible spectrum. This is a not surprising result, since those networks were not trained for such a task, however it tell us that we cannot use this trained networks without modifications (*fine-tunning*) on cross-spectral applications. Moreover, the 2ch-country network outperforms all the other networks and descriptors in all the categories by a surprising margin. Clearly, as pointed out in [59], the key on the performance of the 2ch network is that the information is jointly processed right from the first layer.

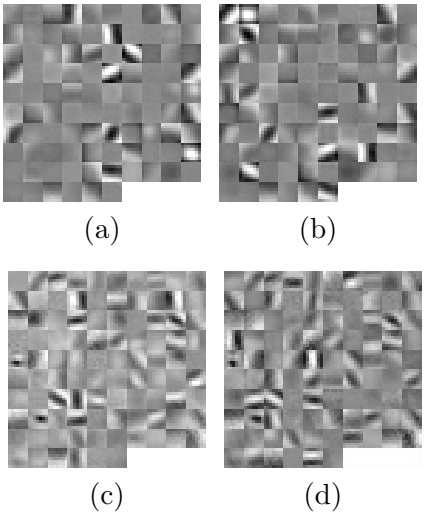


Figure 5.4: Visualization of the first layer filters of: (a) and (b) 2ch network filters trained in the visible spectrum domain (Yosemite); (c) and (d) 2ch network filters trained in the VIS-NIR cross-spectral domain (our best case).

A visual comparison of the first layer filters learned by the networks can be seen on Figure 5.4. Here we can see that our best model has learned similar filters to those presented in (a) and (b); somehow this means that the first layer features learned for image matching in different spectra are quite similar to those from grayscale image matching. We can also see from the filters that our trained network searches for lines and edges rather than textures, information that can be lost by switching to a different spectrum. This is interesting, since having similar first layer filters means that fine-tunning techniques can be applied to VIS similarity networks to work in cross-spectral domains. However, the success of these techniques will depend on how similar are the base datasets [58].

Descriptor/Network	Fi	Fo	In	Mo	Ol	St	Ur	Wa	Mean
SIFT [33]	39.4	11.3	10.1	28.6	19.6	31.1	10.8	40.3	23.9
GISIFT [14]	34.7	16.6	10.6	19.5	12.5	21.8	7.2	25.7	18.6
EHD [1]	48.6	23.17	30.2	33.9	19.6	27.3	3.7	23.5	26.6
LGHD [2]	18.8	3.7	8.1	11.3	8.2	6.7	7.4	13.9	9.8
2ch liberty [59]	30.1	1.8	4.5	24.1	8.2	15.2	2.2	35.8	17.2
2ch notredame [59]	26.7	1.7	4.6	21.4	9.0	15.9	2.9	33.0	16.4
2ch yosemite [59]	36.3	1.7	5.5	30.7	12.6	17.2	4.3	38.6	20.6
siam liberty [59]	38.4	27.0	19.1	27.7	16.5	26.0	12.0	31.8	26.1
siam notredame [59]	36.2	25.6	13.3	24.4	16.7	25.2	11.6	30.0	24.2
siam yosemite [59]	33.7	20.8	20.8	22.2	18.7	21.5	17.2	27.7	23.6
2ch-country	**9.9**	**0.1**	**4.4**	**8.8**	**2.3**	**2.1**	**1.5**	**6.4**	**4.4**
siam-country	15.7	10.7	11.6	11.1	5.2	7.5	4.6	10.2	9.6
psiam-country	17.0	9.8	11.1	11.8	6.7	8.2	5.6	12.0	10.3
PN-Net RGB	25.3	4.3	7.0	19.4	7.3	10.2	5.0	17.8	12.05
PN-Net NIR	24.7	4.6	6.5	15.8	7.8	10.8	4.7	16.5	11.40
PN-Net RANDOM	24.6	3.9	6.6	16.0	6.8	9.5	4.4	15.6	10.9

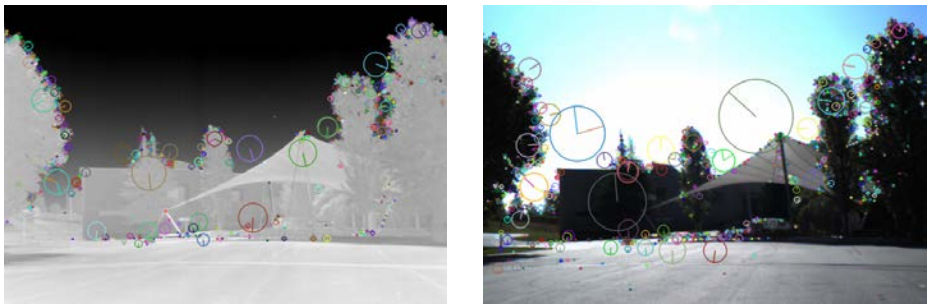
Table 5.4: Performance on the VIS-NIR local image patches dataset. The results correspond to the false positive rate at 95% Recall (FPR95). The smallest the better. The remaining eight categories from the dataset presented in [15] are referred to as: Fi=Field, Fo=Forest, In=Indoor, Mo=Mountain, Ol=Old-building, St=Street, Ur=Urban and Wa=Water.

5.4.2 VIS-LWIR image matching

The proposed network has been also evaluated as replacement of local feature descriptors, i.e., we detect local feature points in each image pair, we extract patches of 64x64 around each feature point and then we do the matching using our trained networks in a brute-force manner. To that purpose, we selected the public VIS-LWIR cross-spectral dataset from Chapter 3.

The selection of the local feature detector to be used in this evaluation was not an easy task. In general local features detected in images from different spectrum are different—a kind of low repeatability (see Figure 5.5). Hence, to minimize this inherent drawback of working with cross-spectral images, we end up using custom FAST [46] settings in each image pairs to have a similar response in both spectra. These custom FAST settings increase the number of correct correspondences in the VIS-LWIR cross-spectral scenario; as already mentioned, cross-spectral feature point detection is still an open problem that needs special care in the tuning of user defined parameters.

We evaluate the performance of our networks using the mean average precision (mAP) as in the well-known local feature descriptor benchmark from [36], where the average precision corresponds to the area under the precision-recall curve—recall 1 correspond to the best possible result. Figure 5.6 shows the results of our evaluation.



(a) Thermal (b) Visible

Figure 5.5: Visualization of cross-spectral feature detection using SIFT. The left is a LWIR image and the right the corresponding VIS image. **This figure is best viewed in color**.

Similar to the VIS-NIR case the 2ch-country network outperformed all the other networks and cross-spectral descriptors. On the contrary, 2ch networks trained in the visible spectrum did not perform better than SIFT. Moreover, the results show that a 2ch network model trained in the VIS-NIR cross-spectral scenario can obtain a high mAP when matching image pairs from the VIS-LWIR domain. This generalization is mainly because in both scenarios some of the same problems persist, like: loss of texture and differences in the gradient directions. The generalization capability is an important fact since the number of images from the VIS-NIR spectra available is considerably higher in comparison with those from the VIS-LWIR spectra.

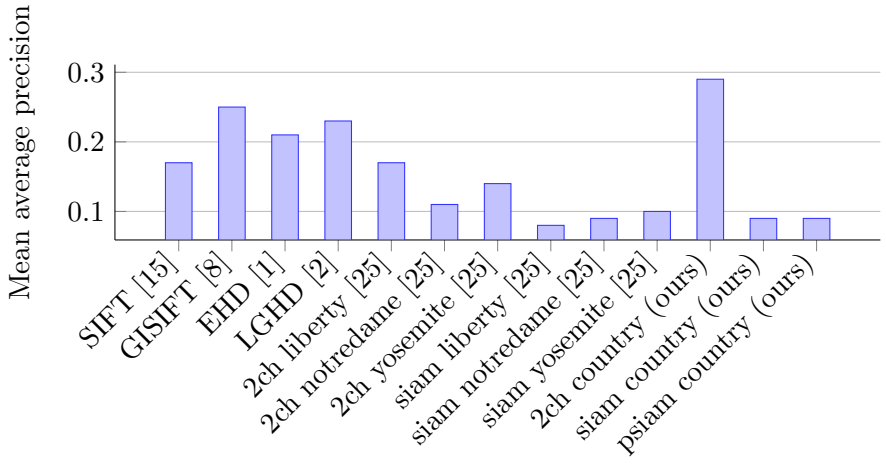


Figure 5.6: VIS-LWIR local feature descriptor performance. The results correspond to the mean average precision over all the images in the dataset. The bigger the better.

5.5 Conclusions

The cross-spectral similarity measure is a challenging task. Our results show that using CNNs to determine the similarity between two patches from different spectra is feasible, and more important it outperforms other alternatives. As an interesting conclusion, in our experiments, a network trained on a VIS-NIR cross-spectral dataset has been later on used in a VIS-LWIR dataset, outperforming the state-of-art in cross-spectral image descriptors. This is an important result since the amount of public data available in the LWIR spectrum is smaller than in other spectra.

Chapter 6

Learning cross-spectral feature descriptors with a quadruplet network

In this chapter we present a novel CNN-based architecture, named Q-Net, to learn how to describe image patches from two different spectral bands in the same way. Given correctly matched and non-matching training image pairs, we train a quadruplet network to map input image patches to a common vectorial representation. Our method is inspired by the triplet network presented in the previous chapter but adapted for cross-spectral scenarios, where, for each image patch, we have at least two possible matching samples, one per each spectrum.

6.1 Introduction

In the previous chapter we tested and evaluated different network architectures that can be used to describe image patches. Results showed that CNN-based methods, designed to work with visible monocular images, could be successfully applied to the description of images from two different spectral bands, with just minor modifications. Even more, results showed that such methods improved the state-of-the-art performance in two cross-spectral benchmarks, outperforming previous hand-made solutions.

In this chapter we present a novel CNN-based network model that is specifically intended to describe image patches from two different spectral bands. Figure 6.1 shows an illustration of our proposal, referred to as Q-Net. Similarly to the PN-Net, described in the previous chapter, Q-Net is mostly a descriptor training network that consists of four copies of the same convolutional neural network, i.e., weights are shared between all CNN, which accepts as input two matching pairs from different spectral bands. Once the network is feed by with the input patches, Q-Net

computes several L_2 distances between the resulting outputs of each CNN, obtaining the hardest cases of matching and non-matching pairs. In other words, the matching pair with the biggest L_2 distance, and the non-matching pair with the smallest L_2 distance, which later are used during the backpropagation process; other pairs are not backpropagated. This is a type of data mining, which at a training time always uses the hardest matching and non-matching pairs to feed the network. At testing, our network behaves as a classic feature descriptor, just needing one of the four CNN to work. Thus, our model is drop-in replacement to hand-made feature descriptors.

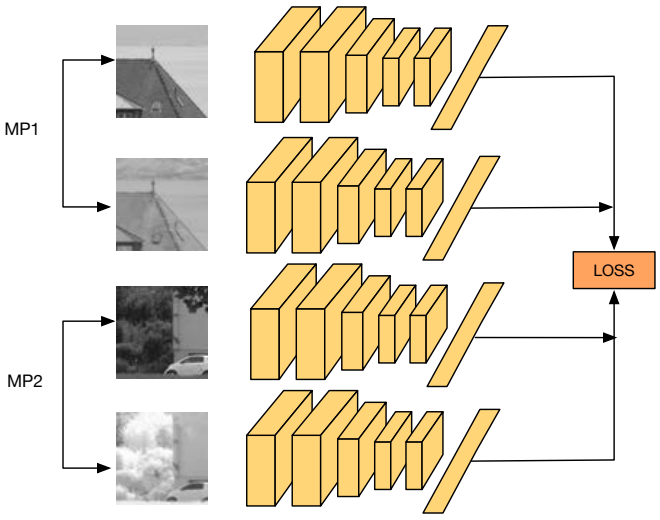


Figure 6.1: Q-Net consists of four copies of the same CNN that accepts as input two different cross-spectral correctly matched image pairs (MP1 and MP2). The network computes the loss based on multiples L_2 distance comparisons between the output of each CNN, looking for the matching pair with biggest L_2 distance and the non-matching pair with the smallest L_2 distance. Both cases are then used for backpropagation of the network. This can be seen as positive and negative mining.

Our work is based on the recent success of the triplet network presented in [4] but adapted to work with cross-spectral image pairs, where for each matching pair, there are two possible non-matching patches; one for each spectrum.

6.2 Proposed approach

The motivations behind our quadruplet network are straightforward. As stated before, for each cross-spectral matching pair we have at least two non-matching patches from another spatial location, each one from one of the spectra to be trained. Similar to triplets, we propose Q-Net, a quadruplet network for learning cross-spectral feature

descriptors.

6.2.1 Network architecture

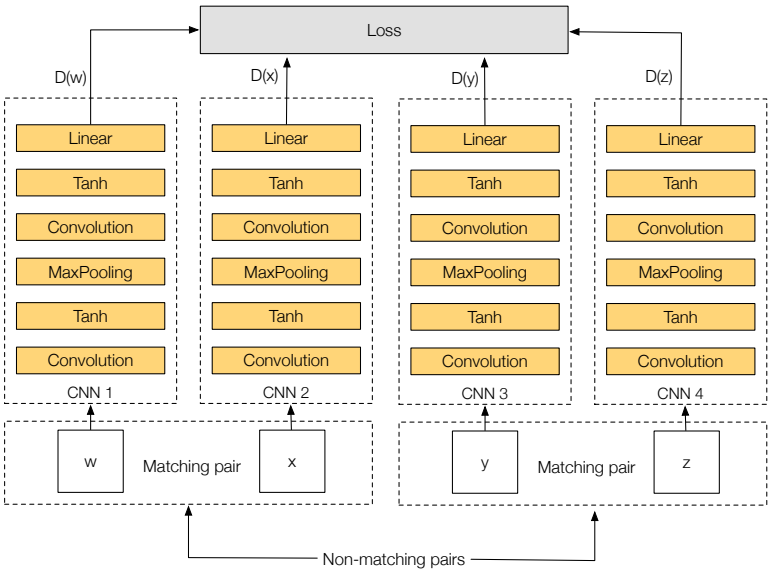


Figure 6.2: Q-Net training quadruplet architecture.

The architecture of Q-Net is similar to PN-Net, but using four copies of the same network instead of three (see Figure 6.2). The input is a tuple Q, with four different input patches $Q = \{w, x, y, z\}$, that is formed by two different cross-spectral matching pairs: (w, x), and (y, z), allowing the network to mine not just non-matching cross-spectral image pairs at each iteration, but also cross-spectral correctly matched pairs.

6.2.2 Loss function

Q-Net loss function extends the mining strategy from PN-Net presented in Chapter 5. Specifically, we add two more distance comparisons to P_{nm}, making the loss suitable for cross-spectral scenarios, and we extend the mining strategy from the non-matching pairs to the correctly matched pairs. At each training step, the network uses the matching pair with larger L_2 distance and the non-matching pair with the smallest L_2 distance. The loss function is defined as follows:

$$P_m = \frac{e^{max(\Delta_1^+, \Delta_2^+)}}{e^{min(\Delta_1^-, \Delta_2^-, \Delta_3^-, \Delta_4^-)} + e^{max(\Delta_1^+, \Delta_2^+)}} \tag{6.1}$$

$$P_{nm} = \frac{e^{min(\Delta_1^-, \Delta_2^-, \Delta_3^-, \Delta_4^-)}}{e^{min(\Delta_1^-, \Delta_2^-, \Delta_3^-, \Delta_4^-)} + e^{max(\Delta_1^+, \Delta_2^+)}} \tag{6.2}$$

$$Loss(Q_i) = P_m^2 + (P_{nm} - 1)^2 \tag{6.3}$$

where, Δ_1^+ corresponds to the L_2 distance between the descriptors of the matching pair w and x, $||D(w), D(x)||_2$; Δ_2^+ corresponds to the L_2 distance between the descriptors of the matching pair y and z, $||D(y), D(z)||_2$; Δ_1^- corresponds to the L_2 distance between the descriptors of the non-matching pair w and y, $||D(w), D(y)||_2$; Δ_2^- corresponds to the L_2 distance between the descriptors of the non-matching pair x and y, $||D(x), D(y)||_2$; Δ_3^- corresponds to the L_2 distance between the descriptors of the non-matching pair w and z, $||D(w), D(z)||_2$; and Δ_4^- corresponds to the L_2 distance between the descriptors of the non-matching pair x and z, $||D(x), D(z)||_2$.

The proposed loss function takes into account all the possible non-matching combinations. For example, if we want to train a network to learn similarities between the VIS and the NIR spectral bands, P_{nm} will compare two VIS-NIR non-matching pairs, one VIS-VIS non-matching pair and one NIR-NIR non-matching pair; instead of using a random function as we did with PN-Net proposed in the previous Chapter. Moreover, since we are trying to learn a common representation between the NIR and the VIS, comparing VIS-VIS and NIR-NIR cases helps the network to have more training examples. Since it is necessary to have two cross-spectral matching pairs to compute P_{nm}, it was natural to extend the mining strategy to P_m, obtaining at each step the cross-spectral matching pair with the larger L_2 distance.

Our method allows learning cross-spectral distances, mining positives and negatives samples at the same time. This approach can also be used in mono-spectral scenarios, providing a more efficient mining strategy than previous works. Results that support our claim are presented in the next section. More importantly, our method can be extended to other cross-spectral or cross-modality scenarios. Even more, it can be extended to other applications such as heterogeneous face recognition, where it is necessary to learn distance metrics between faces from different spectral bands.

6.2.3 Training

Q-Net was trained using Stochastic Gradient Descent (SGD) with a learning rate of 1.1, weight decay of 0.0001, batch size of 128, momentum of 0.9 and learning rate decay of 0.000001. Trained data were shuffled at the beginning of each epoch and each input patch was normalized to its mean intensity. The trained data was split up into two, where 95% of the data was used as training data and 5% as validation. Training was performed with and without data augmentation (DA), were the augmented data were obtained by flipping the images vertically and horizontally, and rotating the images by 90, 180 and 270 degrees. Each network was trained ten times to account

for randomization effects in the initialization. Lastly, we used a grid search strategy to find the best parameters.

Network layer details are described in Table 6.1. The layers and parameters are the same from [4], which after several experimental results showed to be suitable for describing cross-spectral patches. Notice that for feature description shallow models are suitable, since lower layers are more general than the upper ones.

All the code was implemented using the Torch framework ([12]). The GPU consisted of an NVIDIA Titan X and the network was trained in between five and ten hours when we used data augmentation.

Layer	Description	Kernel	Output Dim
1	Convolution	7x7	32x26x26
2	Tanh	-	32x26x26
3	MaxPooling	2x2	32x13x13
4	Convolution	6x6	64x8x8
5	Tanh	-	64x8x8
6	Linear	-	256

Table 6.1: Q-Net layer descriptions.

6.3 Experimental evaluation

6.3.1 Cross-spectral patch matching

In this section we evaluate the performance of our network on the VIS-NIR dataset presented in Chapter 3. As in [3], we train on the country sequence and test in the remaining eight categories.

Results are shown in Table 6.2. Firstly, we evaluated EHD ([1]) and LGHD ([2]), two *hand-made* descriptors that were used as a baseline in terms of matching performance. The performance of LGHD is under 10% and can be considered as state-of-art results—before the current work. Secondly, we test a siamese L_2 network based on the work of [59] that performs better than EHD, but worst than the state-of-art. Thirdly, PN-Net and its variant were tested, not being able to surpass the performance of LGHD without using data augmentation. On the other case, Q-Net showed to be better than the state-of-art even without data augmentation, showing the importance of mining on the non-matching and matching samples in cross-spectral scenarios. Additionally, we tested our model increasing the training data using the previously detailed data augmentation technique, improving the state-of-the-art by a 2.91%. For a more detailed comparison of the different feature descriptors evaluated in the current work, we provide the corresponding ROC curves in Fig. 6.3.

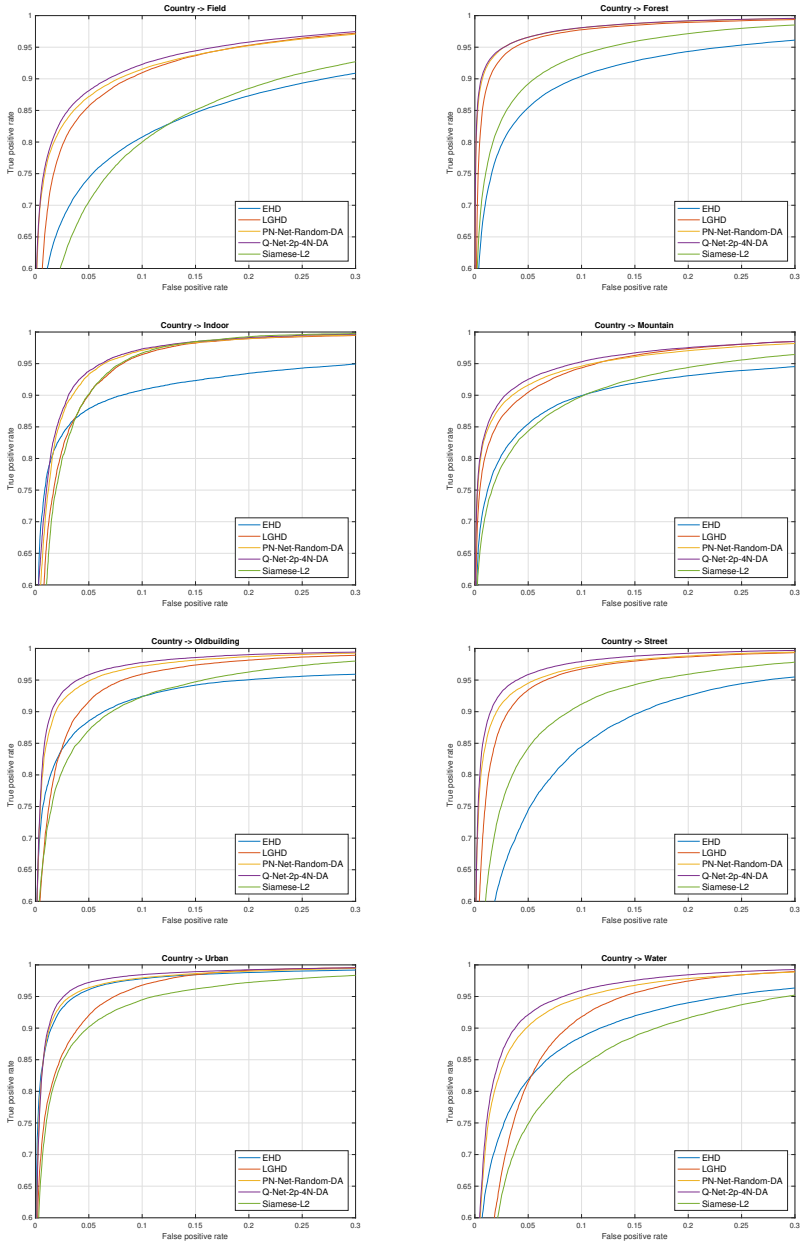


Figure 6.3: ROC curves for the different descriptors evaluated on the VIS-NIR dataset. For Q-Net and PN-Net we selected the network with the best performance. **This figure is best viewed in color.**

Descriptor/Network	Fi	Fo	In	Mo	Ol	St	Ur	Wa	Mean
EHD	48.62	23.17	30.25	33.94	19.62	27.29	3.72	23.46	26.26
LGHD	18.80	3.73	8.16	11.34	8.17	6.66	7.39	13.90	9.77
Siamese-L2	38.47	12.46	7.94	22.36	15.70	16.85	11.06	29.18	15.50
PN-Net RGB	25.33 (1.08)	4.41 (0.28)	7.00 (0.32)	19.37 (1.07)	7.31 (0.32)	10.21 (0.46)	5.00 (0.27)	17.79 (0.67)	12.05 (0.40)
PN-Net NIR	24.74 (0.98)	4.45 (0.14)	6.54 (0.25)	15.75 (0.44)	7.78 (0.19)	10.82 (0.25)	4.66 (0.14)	16.49 (0.34)	11.40 (0.15)
PN-Net Random	24.56 (1.00)	3.91 (0.20)	6.56 (0.43)	15.99 (0.60)	6.84 (0.31)	9.51 (0.36)	4.407 (0.34)	15.62 (0.61)	10.92 (0.34)
Q-Net 2P-4N	20.80 (0.81)	3.12 (0.20)	**6.11** (0.27)	12.32 (0.49)	5.42 (0.13)	6.57 (0.40)	3.30 (0.11)	11.24 (0.50)	8.61 (0.14)
PN-Net Random DA	20.09 (0.65)	3.27 (0.27)	6.36 (0.14)	11.53 (0.57)	5.19 (0.20)	5.62 (0.20)	3.31 (0.28)	10.72 (0.36)	8.26 (0.24)
Q-Net 2P-4N DA	**17.01** (0.33)	**2.70** (0.17)	6.16 (0.18)	**9.61** (0.38)	**4.61** (0.18)	**3.99** (0.09)	**2.83** (0.13)	**8.44** (0.14)	**6.86** (0.09)

Table 6.2: FPR95 performance on the VIS-NIR scene dataset. Each network, i.e., siamese-L2, PN-Net and Q-Net, were trained in the country sequence and tested in the other sequences as in [3]. Smaller results indicate better performance. In brackets the standard deviation is provided. The remaining eight categories from the dataset presented in [15] are referred to as: Fi=Field, Fo=Forest, In=Indoor, Mo=Mountain, Ol=Old-building, St=Street, Ur=Urban and Wa=Water.

6.3.2 Monocular patch matching

Although the proposed approach has been motivated to tackle the cross-spectral problem, in this section we evaluate the proposed architecture when just a visible spectrum dataset is considered. This is intended to evaluate the validity of the proposed approach in classical scenarios.

For the evaluation we used the *multi-view stereo correspondence dataset* from [57], which is considered a standard benchmark for testing local feature descriptors in the visible domain (e.g., [4,23,50,59]). The dataset contains more than 1.2 million patches of size 64x64 divided into three different sets: Liberty, Notredame and Yosemite, where each image patch was computed from *Difference of Gaussian* (DOG) maxima. We followed the standard protocol of evaluation, training our network three times, one at each sequence, and testing the FPR95 in the remaining two sequences. In our evaluation, we compared our model against two other learned L_2 descriptors, the first from [59] and the second from [4]; which can be considered state-of-the-art.

Quadruplets networks were trained using Stochastic Gradient Descent (SGD) with a learning rate of 0.1, weight decay of 0.0001, batch size of 128, momentum of 0.9 and learning rate decay of 0.000001. Trained data was shuffled at the beginning of each epoch and each input patch was normalized using zero-mean and unit variance. We split up each training sequence into two sets, where 80% of the data were used as training data and the 20% left as validation data. We used the same software and hardware from the previous experiment. As in the previous experiment, Q-Net and PN-Net networks were trained ten times to account for randomization effects in the

Descriptor		Siamese-L2	PN-Net-1	PN-Net-2	Q-Net
Training	**Testing**				
Notredame	Yosemite	15.2	8.5	8.5	**7.7**
	Liberty	12.5	9.2	8.9	**7.6**
Yosemite	Notredame	18.8	4.5	4.4	**4.1**
	Liberty	8.4	10.8	10.8	**10.2**
Liberty	Yosemite	20.1	9.5	**8.8**	9.3
	Notredame	6.0	4.1	4.0	**3.8**
mean		13.5	7.8	7.5	**7.1**

Table 6.3: Matching results in the *multi-view stereo correspondence dataset*. Evaluations were made on the 100K image pairs groundtruth recommended from the authors. Results correspond to FPR95. Smallest results indicate better performance. In brackets the standard deviation is provided. PN-Net-1 uses 2.560.000 patches, PN-Net-2 3.840.000, and Q-Net 2.560.000.

initialization. This dataset it was firstly introduced in Chapter 3.

Table 6.3 shows the results of our experiments. Q-Net and PN-Net performed better than the Siamese-L2 network proposed by [59], which is an expected result, since the siamese-L2 network was not optimized for L_2 comparison during training as the other two networks were. Q-Net performed better than PN-Net by a small margin but using much less training data. When comparing both techniques with the same amount of data, the difference becomes bigger. Meaning that our network needs less data to train than PN-Net, i.e., Q-Net needs less training data than PN-Net and it converges more quickly.

Regarding training time, both networks perform similarly. In our experiments, PN-Net was about 9% faster than Q-Net when both networks where trained with the same amount of patches. In essence, the improved accuracy performance of Q-Net is related to a small loss in training speed.

6.3.3 Network parameters

In addition, we tested the performance when different descriptor sizes were used. Fig. 6.4 shows the results of our experiment. From the figure we can see that there is a gain in increasing the descriptor size until 256. Descriptor sizes bigger than 256 did not perform better.

6.4 Conclusions

Experimental results with a VIS-NIR dataset showed the validity of the proposed approach, improving the state-of-the-art by almost 3%. Additionally, results showed that the proposed approach is also valid for training local feature descriptors in the

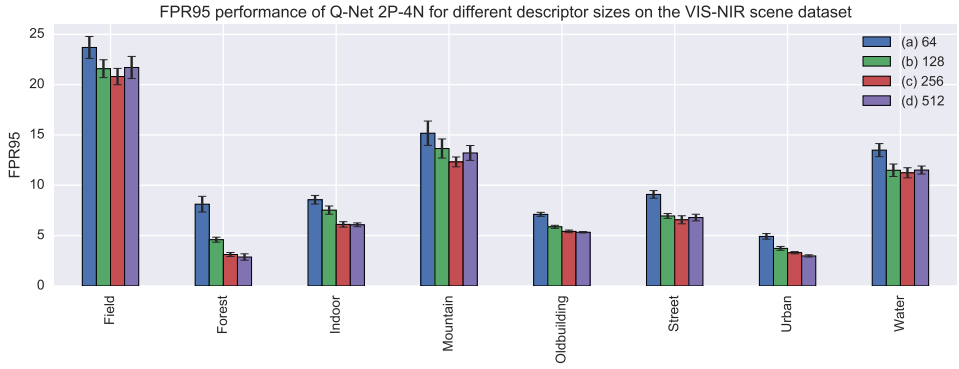


Figure 6.4: FPR95 performance on the VIS-NIR scene dataset for Q-Net 2P-4N using different descriptor sizes ((a) 64, (b) 128, (c) 256 and (d) 512). Shorter bars indicate better performances. On top of the bars standard deviation values are represented with segments.

visible spectrum, providing a network with similar performance to the state-of-the-art, but requiring less training data.

Chapter 7

Application of cross-spectral features

In this chapter we explore the application of cross-spectral feature descriptors as part of a visual odometry system for ground vehicles based on the simultaneous utilization of cameras from different spectral bands. It encompasses the stereo rig described previously in Chapter 3, composed of an optical (visible) and a thermal sensors. Log-Gabor wavelets at different orientations and scales are used to extract interest points from both images. These are then described using a combination of frequency and spatial information within the local neighborhood. Matches between the pairs of multimodal images are computed using the cosine similarity function based on the descriptors. Pyramidal Lucas–Kanade tracker is also introduced to tackle temporal feature matching within challenging sequences of the data sets. The vehicle egomotion is computed from the triangulated 3-D points corresponding to the matched features. A windowed version of bundle adjustment incorporating Gauss–Newton optimization is utilized for motion estimation. An outlier removal scheme is also included within the framework to deal with outliers.

7.1 Introduction

During the last decade, the automotive industry witnessed the introduction of a variety of cameras to enhance vehicles's safety (e.g., thermal and parking cameras). Developing localization techniques on these grounds represents an interesting research path for the coming years. Studies on using visual information for self-localization have been conducted over the last decades. Visual odometry (VO) along with visual simultaneous localization and mapping (SLAM) represent the main vision driven localization solutions. VO involves the estimation of the egomotion of an agent using only visual information from one or multiple cameras. It has been widely investigated in computer vision and robotics. Early attempts to recover motion from vision were

made as far as three decades ago [37]. VO was coined as so for the first time in [42]. Its applications span a variety of domains such as robotics, automotive, and space missions. In the context of driving assistance and autonomous systems, self-localization represents a fundamental issue. The vehicle's own movement (egomotion) is a prerequisite for higher level tasks (e.g., scene perception). In general, this task is performed using wheel odometry, Inertial Measurement Units (IMUs), or GPS devices. Another way to accomplish that task is through VO, which takes advantage of information from cameras. This information can overcome negative aspects of wheel odometry, particularly in slippery terrain. In addition, cameras can mitigate the drawbacks of IMUs by providing less drifty estimates of the motion. Furthermore, GPS devices, although very costly, can suffer shortages or inaccuracies. In this case also VO comes as a cheaper and reliable alternative.

In this Chapter, the feasibility of egomotion estimation from cameras working in different spectral bands is explored. The aim is to extend the concepts of VO to multispectral odometry (MO). In the field of driving assistance systems, these types of cameras are already deployed to tackle a variety of problems. Infrared (IR) cameras are used to improve night-time driving experience as they are able to capture scene elements in the dark. Pedestrian detection and collision avoidance mechanisms based on day-time cameras were extended to night-time using IR technology. Our motivation is to take advantage of equipment already in place to get more functionalities. The vehicle motion is estimated incrementally on a frame-to-frame basis using only the acquired stereo image pairs with no prior knowledge of the environment. The system is capable of estimating its 6 degrees of freedom (DOF) without use of filtering techniques. These are generally used with SLAM algorithms, where the choice of the filter influences the accuracy of the motion estimates.

7.2 Feature extraction and matching

As discussed in previous Chapters, feature extraction is a low-level image processing task that represents a prerequisite for most computer vision applications. This is particularly true in the case of autonomous navigation applications, where essential information contained within an image needs to be extracted.

7.2.1 Feature extraction

Phase congruency (PC), the adopted feature detector, is derived from the work done by Morrone and Owens [39] based on the local energy model (LEM). This model was shown to successfully explain a number of psychophysical effects in human feature perception [40]. The LEM assumes that image features are located in the frequency domain, where their Fourier components are maximally in phase. Traditionally, intensity-based extractors assume them to be at points of maximal intensity gradients. These classical operators exhibit a common behavior. The corner response varies considerably with image contrast and changes in lighting conditions making

the setting of appropriate thresholds a difficult task. In [27], Kovesi represented the PC at a position x as follows:

$$PC_2(x) = \frac{\sum_n W(x)\lfloor A_n(x)\Delta\Phi(x) - T\rfloor}{\sum_n A_n(x) + \epsilon} \tag{7.1}$$

where

$$\Delta\Phi(x) = cos(\Phi_n(x) - \bar{\Phi}(x)) - |sin(\Phi_n(x) - \bar{\Phi}(x))| \tag{7.2}$$

In (7.1) and (7.2), $A_n(x)$ and $\Phi(x)$ represent, respectively, the amplitude and phase of the n-th component at position x; $W(x)$ is a factor that weights for frequency spread; $\Delta\Phi(x)$ is the phase deviation; T is the estimated noise influence; and ϵ is a small constant added mainly to avoid division by zero. The symbols $\lfloor\ \rfloor$ denote that the enclosed quantity is equal to itself when its value is positive and zero otherwise. This means that only energy values that exceed the noise level T are taken into account in the result. In (7.2), $\bar{\Phi}$(x) represents the weighted mean phase angle. In practice, the PC is computed using banks of Log-Gabor filters at different frequencies and orientations. Our implementation comprises a set of 24 Log-Gabor filters corresponding to six orientations at four frequencies. They are used to obtain the PC map of the images used to extract edges and corners by calculating the maximum (M) and minimum (m) moments

$$M = \frac{1}{2}(c + a + \sqrt{b^2 + (a - c)^2}) \tag{7.3}$$

$$m = \frac{1}{2}(c - a + \sqrt{b^2 + (a - c)^2}) \tag{7.4}$$

where

$$a = \sum (PC(\theta)cos(\theta))^2 \tag{7.5}$$

$$b = \sum (PC(\theta)cos(\theta))sin(\theta)) \tag{7.6}$$

$$c = \sum (PC(\theta)sin(\theta))^2 \tag{7.7}$$

where $PC(\theta)$ represents the PC value determined at orientation θ and the sum operation is performed for the set of the used orientations. At this stage, a given pixel is labeled as an edge if its maximum moment is large. It is labeled as an corner if, at the same time, its minimum moment is also large. Figure 7.1 shows the resulting edge map for a multispectral image pair. In order to improve the detection and matching cross-spectral approach from [41], nonmaxima suppression, and feature spreading were introduced.

1. Nonmaxima Suppression: Once corners are extracted, a common observation is that they might be clustered. This can possibly add ambiguity when matching those features. One solution to tackle this problem is the usage of nonmaxima suppression. It is used in computer vision applications and more specifically

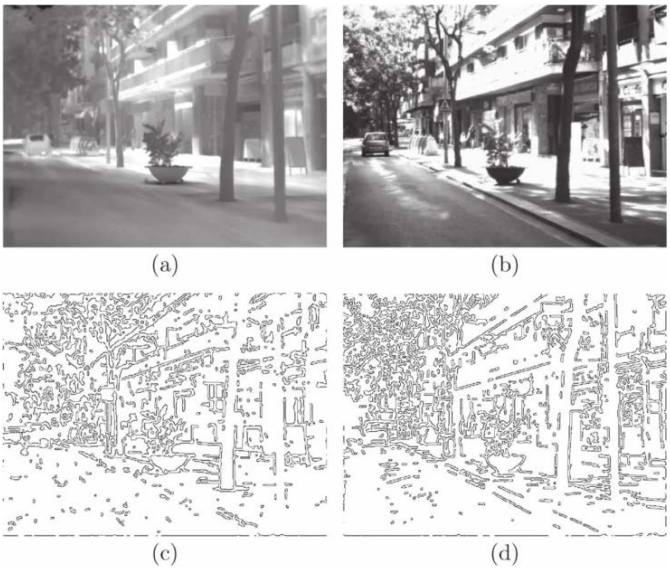


Figure 7.1: IR and visible stereo pair with corresponding edge maps: (a) and (b) IR and visible images; (c) and (d) corresponding edge maps. (Images from our data set).

in feature extraction algorithms [33]. It mainly consists in keeping only corners larger than all their neighbors. Figure 7.2 illustrates the corners obtained before and after applying the nonmaxima suppression using a three-pixel neighborhood.

2. Spreading Features Across the Image: A common problem with feature detectors is that some areas of the images are overloaded with interest points, whereas other regions are left featureless (i.e., nearly empty). This is due to the fact that the detection process is carried out at a small scale where only a restricted area around a given pixel is considered. Fortunately, there are alternatives for this limitation. In our work, the considered image is subdivided into subimages, where the detection takes place. A maximum number of features are allowed per subimage to guarantee the spread of interest points to all image regions if there is enough texture. Figure 7.2 illustrates an example contrasted to the original detection scheme.

7.2.2 Feature description

The next step is matching the extracted keypoints. For this aim descriptors are computed based on the edge histogram descriptor (EHD) [1] and combined with the Log-Gabor coefficients (24 elements) calculated in the previous step. Although this descriptor has lower matching performance than the descriptor proposed in Chapter 4

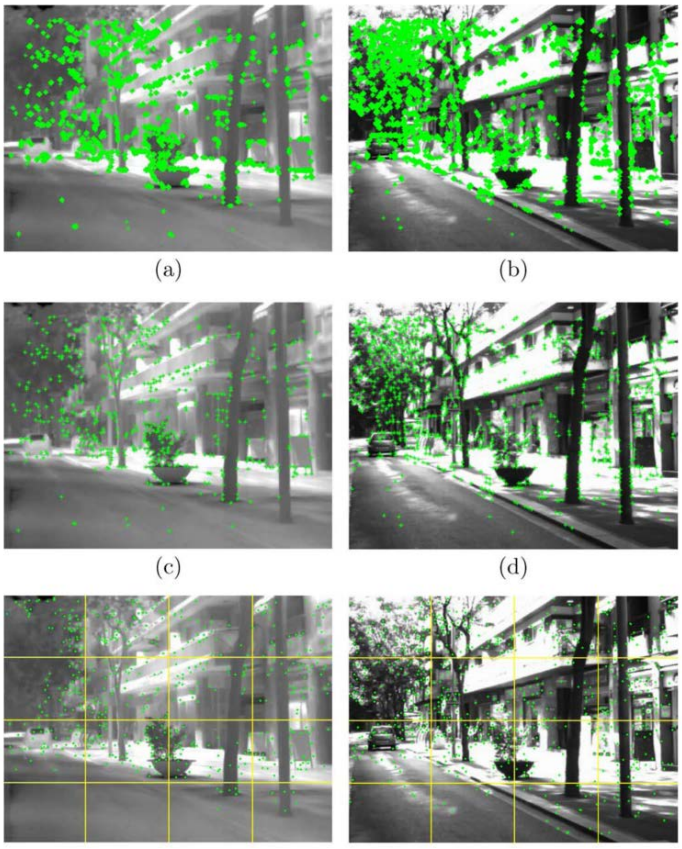


Figure 7.2: Extracted features in a stereo pair (left:IR, right:visible): (a) and (b) Raw; (c) and (d) using nonmaxima suppression; (e) and (f) using subimage extraction.

it is much faster, and thus it fits better for this kind of applications, were the matching speed is important[1].

7.2.3 Matching

There are mainly two types of matching tackled within the scope of this work. This is driven by the fact that at any time t, the algorithm is fed with four input images: left and right at times $t-1$ and t. Therefore, in addition to the stereo matching that takes place every time a stereo image pair is acquired; there is a temporal (sequential) matching that needs to be addressed. For this dual objective, the cosine similarity function is used to compare features descriptors. Let D_L be the descriptor of the

[1]CNN-based solutions were not evaluated, since were introduced after this work

feature f_L at position (x_L, y_L) in the left image. Similarly, let D_R be the descriptor of a potential match f_R at position (x_R, y_R) in the right image within a search window $\text{disp}_x \times \text{disp}_y$ centered at (x_L, y_L). disp_x and disp_y account for the maximum expected horizontal and vertical disparities, respectively. The similarity function is given by

$$S(D_L, D_R) = \frac{\sum d_{L_j} d_{R_j}}{\sqrt{\sum_j d_{L_j}^2 \sum_j d_{R_j}^2}} \tag{7.8}$$

where (D_L, D_R) are the descriptors of the compared features; d_{L_j} , d_{R_j} are, respectively, the j-th coefficients of (D_L, D_R). The feature in the right image that maximizes the similarity function for a given feature in the left image is selected as a potential match. A threshold is then applied to keep only strong matches. As stated above, the algorithm is fed with four images: previous left ($\text{im}L_{t-1}$), previous right ($\text{im}R_{t-1}$), current left (im_{L_t}), and current right (im_{R_t}). The matching is carried out in a loop fashion [20] to keep only features that find their correspondences across all four images. Figure 7.3 illustrates the different steps. We first start by finding stereo matches between ($\text{im}_{L_{t-1}}$) and ($\text{im}_{R_{t-1}}$) (I). Then, sequential matches are found between ($\text{im}R_{t-1}$) and (im_{R_t}) (II). Another stereo matching is performed between (im_{L_t}) and (im_{R_t}) (III). Finally, a last sequential matching is performed between ($\text{im}_{L_{t-1}}$) and (im_{L_t}) (IV). At this stage, if the starting and ending feature points are identical, then the match is accepted. Otherwise, it is simply rejected. This process is carried out for all the features extracted in the first image ($\text{im}_{L_{t-1}}$).

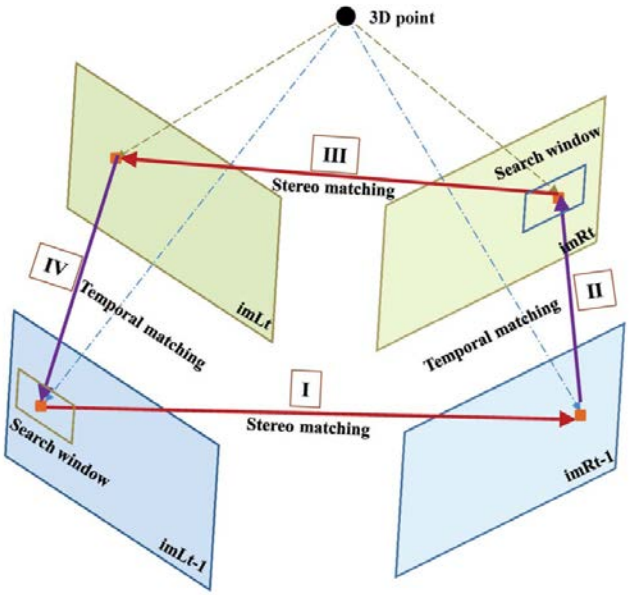


Figure 7.3: Illustration of the loop matching steps.

Since the multispectral image pairs are rectified, the search window (2-D) reduces to a search line (1-D) in the stereo matching process. Correspondences are expected to be found on the same line (i.e., epipolar constraint) of the left and right images. However, this is not the case with the sequential matching where a 2-D search would be still required.

7.3 Motion estimation

The proposed algorithm for egomotion estimation is based on a reduced version of the wide variety of bundle adjustment algorithms surveyed in [54]. This version is called WBA as it analyzes only a portion of the image set to derive the motion estimates. In our case, only the previous and current image pairs of the sequence are used at each time step. First, features are extracted and matched in all four images as described in Section 7.2. Egomotion estimation is achieved using these matches by minimizing reprojection errors using Gauss–Newton optimization within the WBA framework. An outlier rejection scheme based on random sample consensus (RANSAC) [17] is included prior to the final motion optimization step. Outliers that occur due to false matches or matches detected on independently moving objects are dealt with. Each of the aforementioned steps is detailed here along with a reminder of the camera model.

7.3.1 Camera Model

In the current work a multispectral stereo vision setup is considered. The intrinsic and extrinsic calibration parameters of the camera are assumed to be known. Let K be the calibration parameters matrix. Hereinafter, the left camera is considered as the reference camera. The relationship between the homogeneous image coordinates $\hat{x} = (u, v, 1)$ and the camera coordinates $X_C = (X_C, Y_C, Z_C)$ is given by:

$$\hat{x} = K.X_c \tag{7.9}$$

It is worth mentioning that the parameters matrix K is identical for both cameras after rectification of the images. Considering the projections on the left and right images, this yields to:

$$K = K_L = K_R = \begin{pmatrix} \alpha_u & 0 & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{pmatrix} \tag{7.10}$$

where α_u, α_v correspond to the focal length u_0, v_0 the principal point coordinates. Therefore, the projections $\hat{x}_L = (u_L, v_L, 1)$ and $\hat{x}_R = (u_R, v_R, 1)$ on the left and right cameras, respectively, are given by:

$$\hat{x}_L = K - X_C \tag{7.11}$$

$$\hat{x}_R = K.(X_C - (b_L, 0, 0)^T) \tag{7.12}$$

where b_L denotes the stereo baseline. Note that v_L and v_R are identical. It is then convenient to define a vector $y = (u_L, v_L, u_R)$ of the projected coordinates on the stereo images obtained by applying the projection function π to a 3-D point X (with respect to the left camera):

$$y = f(X) = \begin{pmatrix} u_L \\ v_L \\ v_R \end{pmatrix} = \begin{pmatrix} \alpha_u \left(\frac{X}{Z} - u_0\right) \\ \alpha_v \left(\frac{Y}{Z} - v_0\right) \\ \alpha_u \left(\frac{X - b_L}{Z} - u_0\right) \end{pmatrix} \tag{7.13}$$

We assume that the camera parameters do not change with time allowing the bundle adjustment to not recompute them again.

7.3.2 Motion parameters

The camera/vehicle motion can be regarded as a combination of rotations and translations embodied in a motion parameters vector $m = (\phi, \theta, \psi, tx, ty, tz)$. The first three parameters correspond to the Euler rotations and form the rotation matrix $R = (\phi, \theta, \psi)$, whereas the last parameters form the translation vector $t = (tx, ty, tz)$. Writing the transformation matrix $M_p(m)$ derived from the motion parameters gives:

$$M_p(m) = T_{xyz}(t).R_x(\phi).R_y(\theta).R_z(\psi) \tag{7.14}$$

This transformation matrix, in homogeneous coordinates, represents the evolution of the motion of a given vector according to the 6 DOF parameters m. In order to retrieve the motion parameters m, the following bundle adjustment formulation of the reprojection error function is minimized:

$$S(m) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{q} r_j(m, X^{(i)})^2 \tag{7.15}$$

where r_j represents the residuals that are functions of the motion vector m. $X(i)$ corresponds to the observations. By observations, it meant the 3-D coordinates obtained from the triangulation of matched features across a stereo image pair. According to [54], Gauss–Newton optimization postulates that the optimal solution m to (7.15) can be computed in an iterative manner by calculating an increment δm at each iteration using the Jacobian matrix $J \equiv \frac{dr}{dm}$ of the residuals vector with respect to the motion parameters m as:

$$(J^T.J).\delta m = -J^T.r \tag{7.16}$$

where $r \in R_n$ is the residual vector and $(J^T.J)$ represents an approximation of the Hessian matrix [54]. There are typically two reprojection strategies for motion estimation where either points from the previous pair are reprojected into the current frame or the other way round. However, as stated in [45], combining both reprojections yields better estimates of the motion. Following the same strategy, the residuals are defined as $r_d \in \mathbb{R}^6$:

$$r_d = (r_f^T, r_b^T)^T \tag{7.17}$$

where

$$r_f = y_k - \hat{y}_k = y_k - f(M_k(\hat{m}).X_{k+1}) \tag{7.18}$$

$$r_b = y_{k+1} - \hat{y}_{k+1} = y_{k+1} - f(M_k^{-1}(\hat{m}).X_k) \tag{7.19}$$

In (7.18), $\hat{y}_k$ corresponds to the estimated coordinates of the feature on the previous camera frame. Similarly, in (7.19), $\hat{y}_{k+1}$ are the estimated coordinates of the feature on the current frame.

7.3.3 Outlier rejection

In order to improve the accuracy of the motion estimation, the algorithm has to get rid of outliers. They are generally caused by matched features belonging to nonstationary objects or simply undetected false matches from the matching process. One way to deal with outliers is constraining the reprojection error residuals relative to a feature to be bound by a user-defined threshold ϵ. This constraint is expressed by:

$$\left(\sum_{j=1}^{q} r_j(m, X(i)^2 \right) < \epsilon \tag{7.20}$$

To this end, the bundle adjustment estimation is wrapped in a RANSAC scheme. At each iteration three matched points are randomly selected to estimate the motion parameters. The rest of the points are tested and classified as inliers or outliers according to (7.20). The winning solution with the largest number of inliers is then used to refine the motion parameters m.

7.4 Experimental evaluation

We tested the methodology previously described in the multispectral VIS-LWIR visual odometry benchmark introduced in Chapter 3.

The following results are based on both types of sequences namely semiurban (Vid01 & Vid02) and rural (Vid03 & Vid04). Note that the term semiurban is used instead of urban since a good portion of the images contain vegetation. Vid01 represents the simplest scenario, where the vehicle is traveling along a straight road. Vid02 is a more challenging data set within the same environment, which contains speed bumps and bends. Both sequences have proven to be challenging as many non-stationary objects and significant illumination variations were experienced. Vid03 corresponds to a straight road followed by a left bend in a rural environment, where moving vehicles were overtaking our car and where severe lighting conditions were encountered at and after the bend. Vid04 represents a U-turn at a roundabout, where MO suffered from blurred images at the level of the roundabout. The major challenge in this type of scenario is that images lack of nearby features. It is mainly due to the nature of thermal imagery, where thermal response of the road varies less than in the visible spectrum. This causes the corresponding image region to be textureless. The direct impact is an underestimation of the motion as noted in [52]. It is believed to be due to the lack of close-by features combined to the short baseline of the stereo-rig.

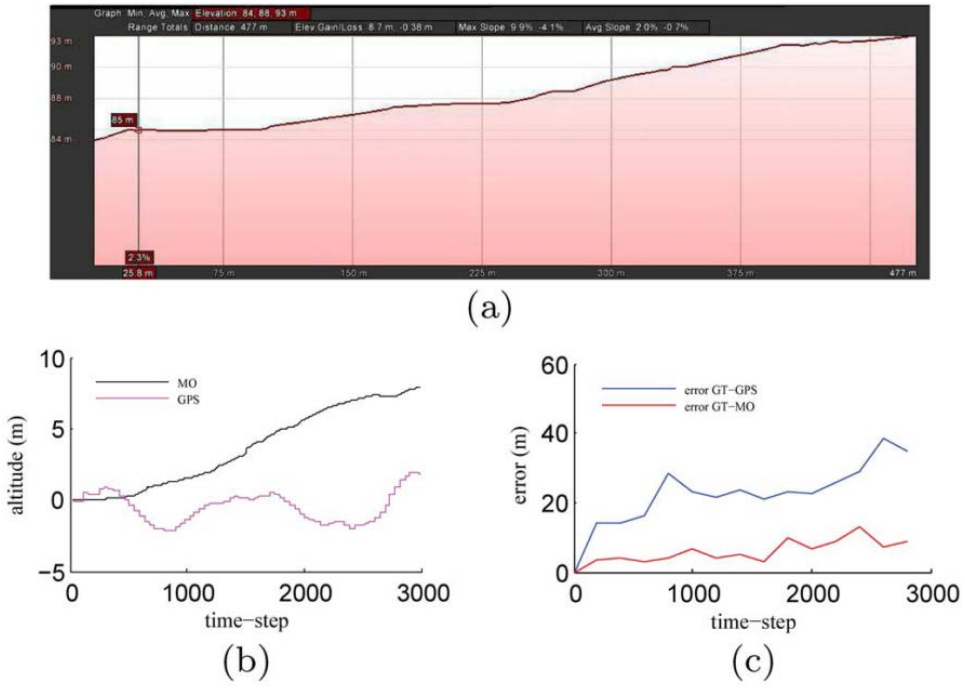


Figure 7.4: Comparison of the MO estimate of the altitude against GPS measurements for Vid01. (a) Google Earth elevation profile of the trajectory. (b) MO estimation of the altitude against GPS measurements. (c) Errors between GT-GPS and GT-MO.

A fixed number of features (1500) is used. This guarantees a reasonable amount

of matches for odometry. In addition, an outlier rejection threshold $\epsilon = 1.5$ is selected. As stated in Chapter 3, the geopositional information is provided by a low-cost GPS considered as 'drifty' ground truth. For this reason, a more precise Google Earth-based ground truth (GT) was manually generated. It was created by introducing control points (based on the images) every 100 frames in Google Earth. This allowed us to obtain a more precise ground truth for comparison with MO estimated trajectories. Figure 7.4(b) illustrates the altitudes estimated by our MO compared with the GPS readings for Vid01. Figure 7.4(a) represents the elevation profile extracted from Google Earth corresponding to the same sequence. It shows that our MO estimates are far more accurate than the GPS measurements. The same observation applies for the estimated trajectory (of the same sequence) as it can be noted in Figure 7.4(c). It illustrates errors between Google Earth-based GT and GPS measurements as well as between GT and MO. Note that errors between GT-GPS are larger than between GT-MO. Following these findings, the same strategy was adopted for all the sequences, where two error graphs are always plotted along with the estimated trajectory. The first error is computed for every frame between MO and the GPS readings that are linearly interpolated due to their low update rate. The second error is the one computed every 100 frames between GT and MO. The errors that are adopted to evaluate the MO performance are based on the GT-MO graphs. Therefore, hereinafter, errors are always expressed from the GT-MO graphs.

Figure 7.5 illustrates the trajectories computed by the GPS and MO for the semiurban sequences (Vid01, Vid02). The peaks in errors (MO-GPS) are due to the imprecision of the GPS and are given as indication only. From Figure 7.5(b) and (d), it can be seen that the achieved results are successful reaching errors as low as 2% and 3% for Vid01 and Vid02, respectively, and defined as:

$$error(\%) = \frac{100.mean(\text{errors})}{\text{traveled distance}} \tag{7.21}$$

These errors do not correspond to the ones commonly provided in the literature and defined as the ratio of the last offset (endpoint) to the traveled distance. The latter errors do not provide information on the behavior of the system along the whole trajectory on the contrary to the errors provided here. The estimated trajectories from the rural sequences are shown in Figure 7.6. Errors obtained in this scenario (rural environment) are slightly higher than in urban sequences for the aforementioned reasons. These errors are shown in Figure 7.6(b) and (d). They correspond to 5% and 4% for Vid03 and Vid04, respectively. In general, the system is able to temporally track 40–50 features due to the textureless nature of thermal imagery. However, in the case of rural scenarios, most of them correspond to far features therefore increasing the errors in the estimation process. In severe lighting conditions, the number of tracked features falls considerably (6–10), making the motion estimation even noisier. Restrictions imposed by the RANSAC-based outlier rejection deal with the wrong matches and allow more robust estimations.

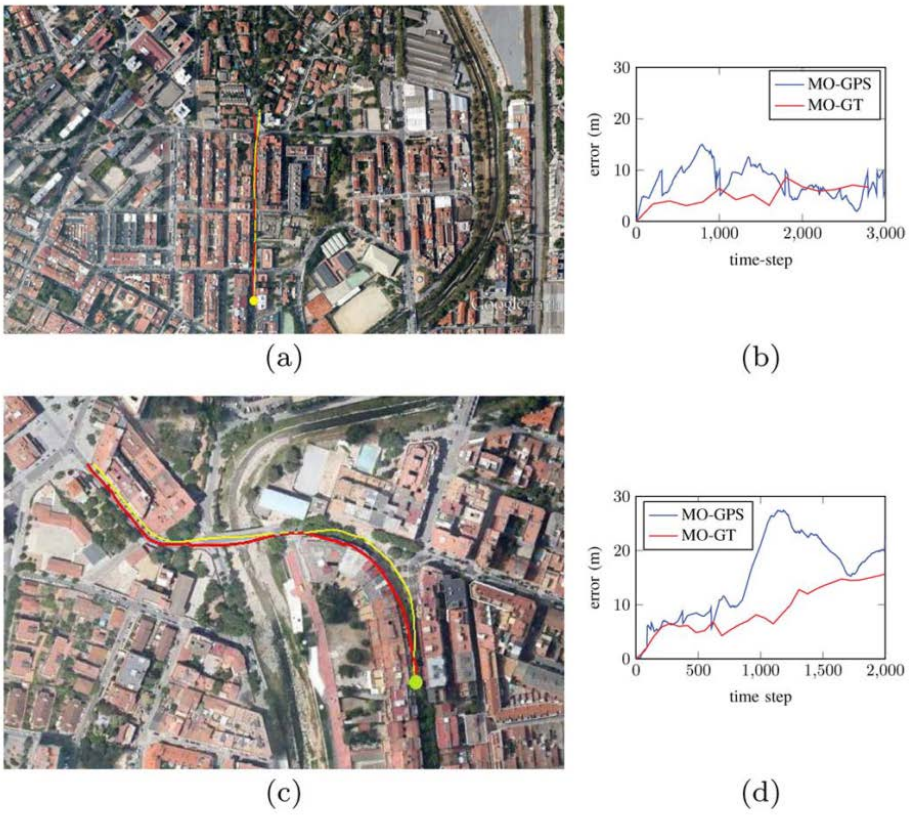


Figure 7.5: MO trajectories and traveled errors for semiurban sequences: (a) and (c) MO trajectories for Vid01 and Vid02 respectively (yellow line: GPS; red line: MO; yellow circle: starting point); (b) and (d) corresponding traveled errors.

7.5 Conclusions

A multispectral stereo odometry solution, that uses cross-spectral matching as a core component, has been introduced To the best of our knowledge, it represents the first attempt in the literature. Features are extracted using a frequency-based detector, namely, PC, and described using a combination of spatial and frequency information. Motion is retrieved using a sliding WBA incorporating Gauss-Newton optimization and RANSAC for outlier removal. Tests were performed under real traffic conditions. Shown results validate our approach and more importantly demonstrate the possibility to achieve MO.

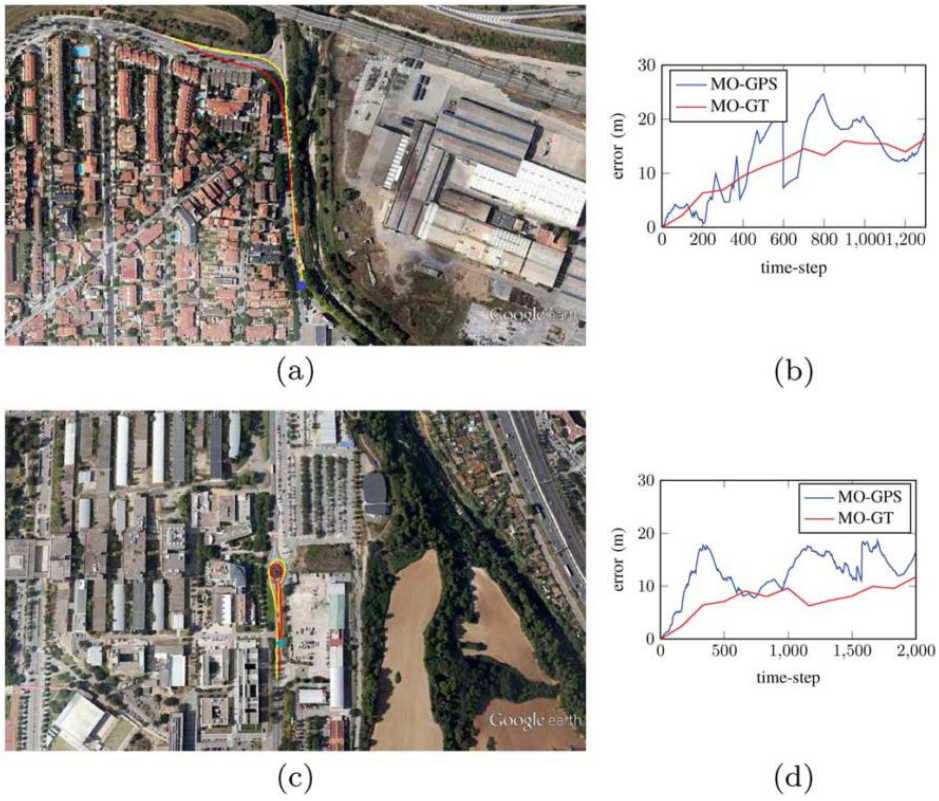


Figure 7.6: MO trajectories and traveled errors for rural sequences: (a) and (c) MO trajectories for Vid03 and Vid04 respectively (yellow line: GPS; red line: MO; blue circle: starting point); (b) and (d) corresponding traveled errors.

Chapter 8

Summary and future work

In this thesis, we review and contribute to the local feature description problem when cross-spectral images are considered. The work starts by proposing a novel approach based on the combination of frequency and spatial information, in a multi-scale scheme, as feature description. Then, we contribute by proposing a CNN based architecture, specifically intended to describe image patches from two different spectral bands. Finally, an application based on the usage of local feature description in cross-spectral domains is presented showing the advantages of working with heterogeneous information. In addition to the three main contributions mentioned above, in this dissertation, two different multi-spectral datasets are generated and shared with the community to be used as benchmarks for further studies. In this chapter, a summary of the thesis and future work are presented.

8.1 Summary

This thesis reviews the state of the art in local feature descriptors and evaluates their performance when they are used in the cross-spectral domain. Coarsely speaking, the thesis contains three sections. The first section includes Chapters 2 and 3, which helps to understand the problem and define the evaluation framework. In Chapter 2 the most relevant work in local feature descriptors, including both classical hand-made and novel CNN based approaches, and cross-spectral applications are reviewed. After this initial study, in Chapter 3 the manuscript presents the benchmarks proposed to evaluate the research work as well as to share with the community working in the multi-spectral domain. It should be mentioned that the algorithms proposed in the whole thesis have been validated with large data sets. In other words, a large amount of work has been devoted to the generation of cross-spectral datasets, which includes the setup of the acquisition systems (acquisition software, external triggers, camera calibration, etc.) and the corresponding image rectification and registration. All the data sets collected during these years in the context of the thesis are now available

for the research community.

The second section includes Chapters 4, 5 and 6. In Chapter 4 the first contribution to the local feature descriptor is given. It consists of a Log-Gabor Histogram Descriptor (LGHD) based on the combined usage of frequency and spatial information. The approach is implemented in a multi-scale and multi-oriented scheme, obtaining better performance than classical hand-made contributions when used in cross-spectral domains. In spite of the obtained results, it should be mentioned that the performance is not as good as in the mono-spectral domain. Hence, to overcome the challenge of working with information from different spectral bands, in Chapters 5 four different CNN based architectures are evaluated to describe image patches. This study helps us to obtain interesting conclusions, such as the possibility to train a network on a VIS-NIR cross-spectral dataset and later on use it in a VIS-LWIR dataset. Additionally, this study inspires the Q-Net architecture proposed in Chapter 6. This novel architecture is specially devoted to tackle the cross-spectral problem. The results obtained with the architecture presented in Chapter 6 overpass the state of the art regarding L_2 CNN-based descriptors; additionally, it shows that this architecture is also useful in mono-spectral domains.

Finally, the last section includes Chapter 7. In this chapter, an application of local feature description using cross-spectral images is presented. It tackles the visual odometry problem showing one useful application of cross-spectral image descriptors.

8.2 Future work

Through the research work during these years different problems have been tackled, and their relationship has been studied. We identified several possibilities to be explored to extend and improve the work presented in this dissertation. The future work comprises short-term challenges and long-term goals as detailed below:

Cross-spectral data set generation: as mentioned in Chapter 3, an essential component of any research work are the evaluation data sets. In the cross-spectral domain there is a limited amount of data sets available to the community. Recently, under the framework of *IEEE Workshop on Perception Beyond the Visible Spectrum*, which is held every year in conjunction with the CVPR conference, a repository collecting public data set is being created. We are working by collecting new cross-spectral data sets (RGB-NIR with new single sensor cameras), which will be ready in the short-term. In the long term we expect to include other spectral bands and generate more elaborated data set by: *i*) segmenting objects in the scene at each spectral band; and *ii*) adding image meta-data (e.g., annotating materials of each object, temperature, etc.).

CNN architectures: One of the biggest challenges in using deep learning solutions in the cross-spectral domain is the low amount of available data. For that reason, an interesting idea to explore in the future is the learning of cross-spectral feature layers, which could be used to train CNN-based solutions in a particular spec-

tral band, and later be used on another, without any modifications. In other words, a network that is trained, knowing the limitation of the other network where is going to be used in the future.

Cross-spectral image based applications: the contributions of this thesis can be applied to other cross-spectral domains. At the moment we are exploring the combined usage of the visible spectrum with ultraviolet, obtaining appealing results in the wood inspection. In the long-term, we expect to do some contribution in the heterogeneous face recognition field using single sensor cross-spectral cameras. Thermal inspection of buildings and applications of NIR imaging in the biological domain will be also explored.

List of Publications

This dissertation has led to the following communications:

Journal Papers

- Cristhian A. Aguilera, Angel D. Sappa, Cristhian Aguilera & Ricardo Toledo. (2017). Cross-spectral local descriptors via quadruplet network. *SENSORS*

- Angel D. Sappa, Cristhian A. Aguilera, Juan A. Carvajal Ayala, Miguel Oliveira, Denis Romero, Boris Vintimilla & Ricardo Toledo. (2016). Monocular visual odometry: A cross-spectral image fusion based approach. *Robotics and Autonomous Systems*.

- Angel D. Sappa, Juan A. Carvajal, Cristhian A. Aguilera, Miguel Oliveira, Dennis Romero & Boris X. Vintimilla. (2016). Wavelet-based visible and infrared image fusion: A comparative study. *SENSORS*

- Tarek Mouats, Nabil Aouf, Angel D. Sappa, Cristhian A. Aguilera & Ricardo Toledo. (2015). Multispectral stereo odometry. *Transactions on Intelligent Transportation Systems*

- Pablo Ricaurte, Carmen Chillán, Cristhian A. Aguilera, Boris X. Vintimilla & Angel D. Sappa. (2014). Feature point descriptors: infrared and visible spectra. *SENSORS*

Conference Contributions

- Cristhian A. Aguilera, Xavier Soria, Angel D. Sappa & Ricardo Toledo. (2017). RGBN multispectral images: a novel color restoration approach. *PAAMS*.

- Cristhian A. Aguilera, Francisco J. Aguilera, Angel D. Sappa, Cristhian Aguilera & Ricardo Toledo. (2016). Learning cross-spectral similarity measures with deep convolutional neural networks. *PBVS, CVPRW*

- Julien Poujol, Cristhian A. Aguilera, Etienne Danos, Boris X. Vintimilla, Ricardo Toledo & Angel D. Sappa. (2016). A visible-thermal fusion based monocular visual odometry. *IBC*
- Mildred Cruz, Boris Vintimilla, Cristhian A. Aguilera, Ricardo Toledo & Angel D. Sappa. (2015). Cross-spectral image registration and fusion: An evaluation study. *EECSS*
- Cristhian A. Aguilera, Angel D. Sappa & Ricardo Toledo. (2015). LGHD: A feature descriptor for matching across non-linear intensity variations. *ICIP*
- Naveen Onkarappa, Cristhian A. Aguilera, Boris X. Vintimilla & Angel D. Sappa. (2014). Cross-spectral correspondences using dense flow fields. *VISSAP*
- Pablo Ricaurte, Carmen Chillán, Cristhian A. Aguilera, Boris X. Vintimilla & Angel D. Sappa. (2014). Performance evaluation of feature point descriptors in the infrared domain. *VISSAP*
- Cristian Duran-Faundez, Cristhian A. Aguilera & Arnoldo S. Norambuena. (2010). Experimenting with RSSI for the perception of moving units in intelligent flexible manufacturing systems. *ICIT*

Bibliography

[1] C. Aguilera, F. Barrera, F. Lumbreras, A. Sappa, and R. Toledo. Multispectral image feature points. *Sensors*, 12(9):12661–72, January 2012.

[2] C. A. Aguilera, A. D. Sappa, and R. Toledo. Lghd: A feature descriptor for matching across non-linear intensity variations. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 178–181, Sept 2015.

[3] Cristhian A. Aguilera, Francisco J. Aguilera, Angel D. Sappa, Cristhian Aguilera, and Ricardo Toledo. Learning cross-spectral similarity measures with deep convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, page 9. IEEE, Jun 2016.

[4] Vassileios Balntas, Edward Johns, Lilian Tang, and Krystian Mikolajczyk. Pn-net: Conjoined triple deep network for learning local image descriptors. *CoRR*, abs/1601.05030, 2016.

[5] F. Barrera, F. Lumbreras, and A. Sappa. Multispectral piecewise planar stereo using manhattan-world assumption. *PRL*, 34(1):52–61, January 2013.

[6] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[7] Luca Bertinetto, Jack Valmadre, João F Henriques, Andrea Vedaldi, and Philip H S Torr. Fully-convolutional siamese networks for object tracking. In *ECCV 2016 Workshops*, pages 850–865, 2016.

[8] Léon Bottou. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade*, pages 421–436. Springer, 2012.

[9] J. Y. Bouguet. Camera calibration toolbox for Matlab, 2008.

[10] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. *BRIEF: Binary Robust Independent Elementary Features*, pages 778–792. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[11] Justyna Cilulko, Paweł Janiszewski, Marek Bogdaszewski, and Eliza Szczygielska. Infrared thermal imaging in studies of wild animals. *European Journal of Wildlife Research*, 59(1):17–23, Feb 2013.

[12] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, number EPFL-CONF-192376, 2011.

[13] Ismael Fernández-Cuevas, Javier Arnáiz Lastras, Víctor Escamilla Galindo, and Pedro Gómez Carmona. *Infrared Thermography for the Detection of Injury in Sports Medicine*, pages 81–109. Springer International Publishing, Cham, 2017.

[14] D. Firmenichy, M. Brown, and S. Süsstrunk. Multispectral interest points for rgb-nir image registration. In *ICIP*, pages 181–184, Brussels, Belgium, September 2011.

[15] D. Firmenichy, M. Brown, and S. Süsstrunk. Multispectral interest points for RGB-NIR image registration. In *ICIP*, pages 181–184, Brussels, Belgium, September 2011.

[16] Philipp Fischer, Alexey Dosovitskiy, and Thomas Brox. Descriptor matching with convolutional neural networks: a comparison to SIFT. *CoRR*, abs/1405.5769, 2014.

[17] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981.

[18] Rikke Gade and Thomas B. Moeslund. Thermal cameras and applications: a survey. *Machine Vision and Applications*, 25(1):245–262, Jan 2014.

[19] Yuan Gao, Boan Pan, Kai Li, and Ting Li. Shed a light in fatigue detection with near-infrared spectroscopy during long-lasting driving, 2016.

[20] A. Geiger, J. Ziegler, and C. Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *2011 IEEE Intelligent Vehicles Symposium (IV)*, pages 963–968, June 2011.

[21] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[22] Alejandro González, Zhijie Fang, Yainuvis Socarras, Joan Serrat, David Vázquez, Jiaolong Xu, and Antonio M. López. Pedestrian detection at day/night time with visible and fir cameras: A comparison. *Sensors*, 16(6), 2016.

[23] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander. C. Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *CVPR*, 2015.

[24] Wenqian Huang, Baihai Zhang, Jiangbo Li, and Chi Zhang. Early detection of bruises on apples using near-infrared hyperspectral image, 2013.

[25] F. Juefei-Xu, D. K. Pal, and M. Savvides. Nir-vis heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 141–150, June 2015.

[26] N. D. Kalka, T. Bourlai, B. Cukic, and L. Hornak. Cross-spectral face recognition in heterogeneous environments: A case study on matching visible to short-wave infrared imagery. In *2011 International Joint Conference on Biometrics (IJCB)*, pages 1–8, Oct 2011.

[27] P. Kovesi. Phase congruency detects corners and edges. In *DICTA*, pages 309–318, Sydney, Australia, December 2003.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[29] Stephen J. Krotosky and Mohan M. Trivedi. Mutual information based registration of multimodal stereo videos for person tracking. *Comput. Vis. Image Underst.*, 106(2-3):270–287, May 2007.

[30] Amanda W. Lewis, Sam T.S. Yuen, and Alan J.R. Smith. Detection of gas leakage from landfills using infrared thermography - applicability and limitations. *Waste Management & Research*, 21(5):436–447, 2003. PMID: 14661891.

[31] José Lezama, Qiang Qiu, and Guillermo Sapiro. Not afraid of the dark: NIR-VIS face recognition via cross-spectral hallucination and low-rank embedding. *CoRR*, abs/1611.06638, 2016.

[32] Ce Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *PAMI*, 33(5):978–994, May 2011.

[33] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[34] Jiayi Ma, Ji Zhao, Yong Ma, and Jinwen Tian. Non-rigid visible and infrared face registration via regularized gaussian fields criterion. *Pattern Recognition*, 48(3):772 – 784, 2015.

[35] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):703–715, Jun 2001.

[36] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1):43–72, Nov 2005.

[37] Hans Peter Moravec. *Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover*. PhD thesis, Stanford, CA, USA, 1980. AAI8024717.

[38] N. J. W. Morris, S. Avidan, W. Matusik, and H. Pfister. Statistics of infrared images. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, June 2007.

[39] M. C. Morrone and R. A. Owens. Feature detection from local energy. *Pattern Recogn. Lett.*, 6(5):303–313, December 1987.

[40] M.C Morrone and R. Owens. Feature detection in human vision: a phase-dependent energy model. *Proceedings of the Royal Society of London B: Biological Sciences*, 235(1280):221–245, 1988.

[41] T. Mouats and N. Aouf. Multimodal stereo correspondence based on phase congruency and edge histogram descriptor. In *Proceedings of the 16th International Conference on Information Fusion*, pages 1981–1987, July 2013.

[42] D. Nister, O. Naroditsky, and J. Bergen. Visual odometry. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–652–I–659 Vol.1, June 2004.

[43] P. Pinggera, T.P. Breckon, and H. Bischof. On cross-spectral stereo matching using dense gradient features. In *Proc. British Machine Vision Conference*, pages 526.1–526.12, September 2012.

[44] Pablo Ricaurte, Carmen Chillan, Cristhian Aguilera-Carrasco, Boris X. Vintimilla, and Angel D. Sappa. Feature point descriptors: Infrared and visible spectra. *Sensors*, 14(2):3690–3701, Feb 2014.

[45] D. Rodriguez and N. Aouf. Robust egomotion for large-scale trajectories. In *2012 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 156–161, Sept 2012.

[46] Edward Rosten, Reid Porter, and Tom Drummond. Faster and better: A machine learning approach to corner detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(1):105–119, January 2010.

[47] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015.

[48] X. Shen, L. Xu, Q. Zhang, and J. Jia. Multi-modal and Multi-spectral Registration for Natural Images. In *ECCV*, pages 309–324, Zurich, Switzerland, Sep 2014.

[49] H. Shengfeng and L. Rynson. Saliency detection with flash and no-flash image pairs. In *ECCV*, pages 110–124, Zurich, Switzerland, Sep 2014.

[50] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015.

[51] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012.

[52] J. P. Tardif, M. George, M. Laverne, A. Kelly, and A. Stentz. A new approach to vision-aided inertial navigation. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4161–4168, Oct 2010.

[53] Wayne Treible, Philip Saponaro, Scott Sorensen, Abhishek Kolagunda, Michael O'Neal, Brian Phelan, Kelly Sherbondy, and Chandra Kambhamettu. Cats: A color and thermal stereo benchmark. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[54] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. *Bundle Adjustment — A Modern Synthesis*, pages 298–372. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000.

[55] Stephen Vidas and Peyman Moghadam. Heatwave : a handheld 3d thermography system for energy auditing. *Energy and Buildings*, 66:445–460, November 2013.

[56] James Z. Wang, Xiaoping Liang, Qizhi Zhang, Laurie L. Fajardo, and Huabei Jiang. Automated breast cancer classification using near-infrared optical tomographic images. *Journal of Biomedical Optics*, 13(4):044001–044001–10, 2008.

[57] S. Winder, G. Hua, and M. Brown. Picking the best daisy. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 178–185, June 2009.

[58] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N.d. Lawrence, and K.q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3320–3328. Curran Associates, Inc., 2014.

[59] S. Zagoruyko and N. Komodakis. Learning to compare image patches via convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4353–4361, June 2015.