



Multimodal Stereo from Thermal Infrared and Visible Spectrum

A dissertation submitted by **José Fernando Barrera Campo** at Universitat Autònoma de Barcelona in partial fulfilment of the requirements for the degree of **Doctor of Philosophy**.

Bellaterra, September 27, 2012

Director	Felipe Lumbreras Universitat Autònoma de Barcelona Dept. Informàtica & Computer Vision Center
Co-director	Angel Sappa Computer Vision Center
Thesis Committee	Dr. Joaquim Salvi Mas Dept. Arquitectura i Tecnologia de Computadors Universitat de Girona Dr. Antonio López Pea Centre de Visió per Computador Dept. Ciències de la Computació Universitat Autònoma de Barcelona Dr. Filiberto Pla Ban Dept. Llenguatges i Sistemes Informàtics Universitat Jaume I Dr. Daniel Ponsa Mussarra Dept. Ciències de la computació Universitat Autònoma de Barcelona Dr. Carme Julià Ferrè Dept. Enginyeria Informàtica i Matemàtiques Universitat Rovira i Virgili



This document was typeset by the author using L^AT_EX 2_ε.

The research described in this book was carried out at the Computer Vision Center, Universitat Autònoma de Barcelona.

Copyright © MMX by José Fernando Barrera Campo. 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 XX-XXXXXX-X-X

Printed by Ediciones Gráficas Rey, S.L.

Abstract

Recent advances in thermal infrared imaging (LWIR) has allowed its use in applications beyond of military domain. Nowadays, this new sensor family is included in diverse technical and scientific applications. They offer features that facilitate tasks, such as detection of pedestrians, hot spots, differences in temperature, among others, which can significantly improve the performance of a system where the persons are expected to play the principal role. For instance, video surveillance applications, monitoring, and pedestrian detection.

During this dissertation is stated the next question: *Could a couple of sensors measuring different bands of the electromagnetic spectrum, as the visible and thermal infrared, provides depth information?* Although is a complex question, we shows that a system of those characteristics is possible as well as their advantages, drawbacks, and potential opportunities.

The fusion and matching of data coming from different sensors, as the emissions registered at visible and infrared band, represents a special challenge, because it has been showed that theses signals are weak correlated. Indeed, they are uncorrelated. Therefore, many traditional techniques of image processing and computer vision are not helpful, requiring adjustments for their correct performs in every modality.

In this research is performed a experimental study that compares different cost functions and matching approaches, in order to build a multimodal stereo system. Furthermore, are identified the common problem between visible/visible and infrared/visible stereo, special in the outdoor scenes. A contribution of this dissertation is the isolation achieved, between the different stage that compose a multimodal stereo system. Our framework summarizes the architecture of a generic stereo algorithm, at different levels: computational, functional, and structural, which is successful because this can be extended toward high-level fusion (semantic) and high-order (prior).

The proposed framework is intended to explore novel multimodal stereo matching approaches, going from sparse to dense representation (both disparity and depth maps). Moreover, context information is added in form of priors and assumptions. Finally, this dissertation shows a promissory way toward the integration of multiple sensors for recovering three-dimensional information.

Resumen

Recientes avances en imágenes térmicas (LWIR) han permitido su uso en aplicaciones más allá del ámbito militar. Actualmente, esta nueva familia de sensor esta siendo incluida en diversas aplicaciones tanto técnicas como científicas. Este tipo de sensores facilitan tareas tales como: detección de peatones, puntos calientes, detección de cambios de temperatura, entre otros. Características que pueden mejorar significativamente el desempeño de un sistema, especialmente cuando hay interacción con humanos. Por ejemplo, aplicaciones de vídeo vigilancia, detección de peatones, análisis de postura.

En esta tesis se plantea entre otras la siguiente pregunta de investigación: *Podría un par de sensores operando en diferentes bandas del espectro electromagnético, como el visible e infrarrojo térmico, proporcionar información de profundidad?* Si bien es una cuestión compleja, nosotros demostramos que un sistema de estas características es posible. Además, de discutir sus posibles ventajas, desventajas y oportunidades potenciales.

La fusión y correspondencia de los datos procedentes de diferentes sensores, como las emisiones registradas en la banda visible e infrarroja, representa un reto atractivo, ya que se ha demostrado que aquellas señales están débilmente correlacionadas. Por lo tanto, muchas técnicas tradicionales de procesamiento de imágenes y visión por computadora son inadecuadas, requiriendo ajustes para su correcto funcionamiento.

En esta investigación se realizó un estudio experimental comparando diferentes funciones de costos multimodal, y técnicas de correspondencia, a fin de construir un sistema estéreo multimodal. También, se identificó el problema común entre estéreo visible/ visible y infrarrojo/visible, particularmente en ambientes al aire libre. Entre las contribuciones de esta tesis se encuentra; el aislamiento de las diferentes etapas que componen un sistema estéreo multimodal. Esta arquitectura es genérica a diferentes niveles, tanto computacional, funcional y estructural, permitiendo su extensión a esquemas mas complejos tales como fusión de alto nivel (semántica) y de orden superior (supuestos).

El enfoque propuesto está destinado a explorar nuevos métodos de correspondencia estéreo, pasando de una solución escasa a una densa (tanto en disparidad como en mapas de profundidad). Además, se ha incluido información de contexto en forma de asunciones y restricciones. Finalmente, esta disertación muestra un promisorio

camino hacia la integración de múltiples sensores.

Contents

Abstract	i
Resumen	iii
1 Multimodal Imagery	1
1.1 Infrared sensors: history, theory and evolution	1
1.1.1 Night vision in ADAS	4
1.2 Multimodal Stereo Head	6
1.3 Multimodal stereo	8
1.3.1 Multimodal stereo head	8
1.3.2 Calibration and rectification	9
1.4 Evaluation Dataset	10
1.4.1 Multispectral datasets	10
2 Similarity Window-based Matching Cost Function	15
2.1 Introduction	15
2.2 Matching Cost Volume	17
2.3 LWIR/VS Matching Cost Functions	18
2.3.1 Mutual Information	19
2.3.2 Gradient Information	22
2.3.3 Mutual and Gradient Information	24
2.3.4 Scale-Space Context	24
2.4 Experiments	25
2.5 Conclusions	28
3 Multimodal Sparse Stereo	29
3.1 Introduction	29
3.2 Matching Cost Volume Computation	32
3.3 Disparity and Depth Computation	33
3.4 Evaluation Methodology	34
3.5 Experiments	36
3.6 Conclusions	42
4 Piecewise Planar Stereo	45
4.1 Introduction	45

4.2	Background	46
4.3	Piecewise Planar Stereo	48
4.3.1	Initial Disparity Map	49
4.3.2	Plane Based Hypotheses	52
4.3.3	Piecewise Planar Labeling	55
4.4	Experiments	57
4.5	Conclusions	63
5	Context-Based Multimodal Stereo	65
5.1	Introduction	65
5.2	Background	67
5.3	The Algorithm	69
5.3.1	Multimodal Matching Cost Volume	70
5.3.2	Plane based Hypotheses Generation	73
5.3.3	Piecewise Superpixel Labeling	78
5.4	Experimental Results	79
5.5	Conclusions	88
6	Conclusions	89
6.1	Summary and contributions of this thesis	89
	Bibliography	91
	Publications	99

Chapter 1

Multimodal Imagery

This section introduces the basic concepts to understand infrared sensors, its evolution, features, and usefulness for ADAS. The discovery of new alloys, materials as well as advances in image processing techniques have allowed the development of new sensors, which work in different spectral bands.

1.1 Infrared sensors: history, theory and evolution

IN 1800 the German astronomer Sir Frederick William Herschel experimented with a new form of electromagnetic radiation, which was called infrared radiation. During his experiments, he built a basic monochromator¹, with which it measured the distribution of energy of a ray of sunlight. The Herschel's experiment demonstrate the existence of radiation beyond what we know as the visible spectrum. Furthermore, provided evidence of a relationship between temperature and color. His experiment used a glass prism to break sunlight up into its constituent spectral colors (Fig. 1.1(a)). Then, a array of thermometers with blackened bulbs were put over every spectral color. In this way, these should measure the temperature of the different colors. During the experiments, Herschel noticed that a thermometer near to experimental setup registered a higher temperature in comparison to the used in the array (visible spectrum). Further experiments confirms that there an invisible form of light beyond the visible spectrum, and it can be measured. This discovery was largely ignored till modern instruments were used to acquire multispectral information, giving rise to a new research field such as is *thermography*.

Clearly, the equipments used by Herschel's have been improved several times. However, nowadays infrared cameras still are based on its operating principle (see Fig. 1.1). Although, infrared radiation is not detectable by the human eye, an IR

¹is an optical device that can produce monochromatic light.

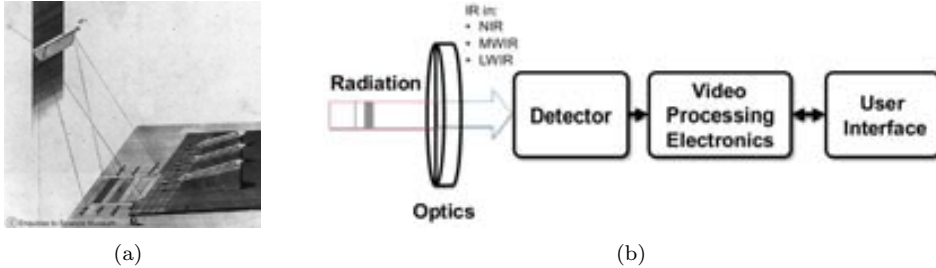


Figure 1.1: (a) Herschel's experimental setup. (b) Simplified block diagram of an IR camera.

cameras can do it. Their operation is similar to a digital camera working on VS however infrared cameras replace the classical Charge Coupled Device (CCD) used in digital cameras by a Focal Plane Array (FPA), which is made of materials and alloys sensitives to IR wavelengths.

Their main components are shown in Fig. 1.1(b), which are a *lens* that focuses infrared radiation onto a *detector*, the *electronics* that convert opto-electrical signals into images, plus a optional software unit that acts as *interface* between the camera and user.

In general, the FPA can be classified into two categories, according to their working principle: (i) photon detectors; and (ii) thermal detectors. The former class corresponds to those detectors where the radiation is absorbed within the material by interaction with electrons. The observed electrical output signal results from the changed electronic energy distribution. These electrical property variations are measured to determine the amount of incidents optical power.

These detectors in turn could be divided into more categories depending on the type of interaction between the FPA and the incident infrared radiation. So, they can be grouped into four main categories: (i) intrinsic detector; (ii) extrinsic detectors; (iii) photoemissive detectors; and (iv) quantum.

They have high performance but require cryogenic cooling. Therefore, IR systems based on semiconductor photodetectors are heavy, expensive and inconvenient to many applications, specially ADAS.

In a *thermal detector* the incident radiation is absorbed by a material semiconductor, causing a change in the temperature of the material, or other physical property, its resultant change is used to generate an electrical output proportional to incident radiation. In this kind of sensor is necessary that at least one inherent electrical property change with temperature, and it could be measured. A traditional device used for thermal detector of low-cost are bolometers, they turn an incoming photon flux into heat, changing the electrical resistance of the detector element, whereas in a pyroelectric detector, for example, this flux changes the internal spontaneous polarization. Currently, thermal detectors are available for commercial applications, opposite to the based on photons, which are restricted to military uses.

In contrast to photon detectors, the thermal detectors do not require cooling. In spite of it, the photon detectors are popularly believed to be rather speed, and selective wavelength in comparison with another type of detectors, this fact was exploited by the military industry. But, later in the '90s, advances in micro miniaturization allowed arrays of bolometers or thermal detectors. It compensated the moderate sensitivity and low frame rate of thermal detectors. Large arrays let high quality imagery and good response time, also the manufacture cost to drop quickly (an extensive review in [72]).

Infrared by definition refers to that part of the electromagnetic spectrum between the visible and microwave region, and their behavior is modeled for the next equations:

$$v = \frac{c}{\lambda}, \quad (1.1)$$

where c is the speed of light, 3×10^{12} (m/sg); v is the frequency (hz), and λ its wavelength (m). The energy is related to wavelength and frequency by the following equation:

$$E = hv = \frac{hc}{\lambda}, \quad (1.2)$$

where, h is Planck constant, equal to 6.6×10^{-34} (joules sg). Notice that, light and electromagnetic waves of any frequency will heat surfaces that absorb them, and the infrared detectors measure the emissivity in this band, but it could occur in other bands, depending on physical properties of the objects (constitutive material). Humans at normal body temperature mainly radiate at wavelengths around $10\mu m$, it corresponds to Long-Wave InfraRed band or LWIR (see the table 1.1).

Table 1.1: General spectral bands based on atmospheric transmission and sensor technology

Spectral Band	Spectral Wavelength (μm)
Visible	0.4 - 0.7
Near InfraRed (NIR)	0.78 - 1.0
Short-Wave InfraRed (SWIR)	1 - 3
Mid-Wave InfraRed (MWIR)	3 - 5
Long-Wave InfraRed (LWIR)	8 - 12

The use of night vision devices should not be confused with thermal imaging. While, night vision devices convert ambient light photons into electrons, which are then amplified by a chemical and electrical process and then converted back into a visible ray of light. The thermal sensors create images detecting radiation that emits the objects.

1.1.1 Night vision in ADAS

Night vision is a technology that was originated in military applications for producing a clear image on the darkest of night. As it was explained above, they need no light whatsoever to operate, and also have the ability to see through special conditions such as: fog, rain, haze, or smoke. Thus, it would be interesting for ADAS since road users can avoid potentially hazards.

Researchers have always been convinced that thermal imaging is an extremely useful technology. Nowadays, it can be found vehicles with IR equipment. This tendency is being followed by different car manufacturer. Then, new technical requirements have been formulated. Today, there are two different technologies on the market: One is called active, using near infrared laser and detectors, and the other passive, which only uses thermal infrared detector [1]. The difference is notorious. Active systems beams infrared radiation into the area in front of the vehicle, for this purpose, usually it involves laser sources or just a light bulb in the near infrared range (NIR). Then, infrared radiation is reflected by objects, the road, humans and other road users. Later, the reflections are captured using a camera sensitive to a same region of the spectrum that was emitted, for example, a NIR camera. Whereas passive systems register relatives differences in heat, or infrared radiation emitted in the far infrared band (FIR), and it does not need a separate light source.

The selection of the best night vision system for ADAS is not easy, different factors must be considered. Although, both systems are technically and economically feasible, the passive systems based on FIR offer advantages. It is not dependent on the power of the infrared beams, because those are not necessary. Then, it contains less components so it is less susceptible to breakdowns. FIR detects people and hot spot at a longer range. The major advantage of FIR is that it is not sensitive to the headlight of oncoming traffic, street lights and powerfully reflecting surfaces such as traffic signs. Since NIR systems (or passives) are based on the use of light beams with wavelength close to visible spectrum, two facts can happen. Firstly, the driver can be blinded for light ray reflection or dispel. Or, if an object is illuminated by two or more infrared beams, this could appear brightly on the screen. The worse case is when an infrared source directly illuminates a detector, situation frequent by the glare of oncoming cars [26, 77].

The setup of IR systems, in the context of ADAS, is another interesting topic to be mentioned. It includes camera position, display, and applications. They are discussed in more detail in next section:

Camera position

The location of the sensor or camera is critical to obtain an acceptable image of the road. If the sensor or camera is positioned low (e.g., in the grill), the perspective of the road will be less than ideal, especially when driving on vertical curves. It is acceptable to position the sensor at the driver eye height, and it is preferable to place it above the driver's eyes. Another aspect of camera position is that a lower position is

more exposed to dirt. Glass interferes with the FIR wavelengths and cannot be placed in front of the sensor. Thus, the FIR sensor cannot be placed behind the windshield. However, early research, as the performed by BMW, concluded that the best position is at the left of the front bumper. This result could be contradictory but a new generation of FIR sensors is being developed, especially for ADAS. Table 1.2 presents the key points of researches performed by five car manufacturer. Other examples are: Renault NIR-contact analogue, which is placed at the inside rear view mirror, and the Daimler-Chrysler camera (NIR), which is placed high above the driver's eyes (rear mirror).

Table 1.2: IR systems.

Manufacturer	Technology and setup	Camera specification
<i>General Motors and Volvo</i>	FIR camera mounted behind the front grill and cover by a protective window.	Raytheon IR-camera. Maximum sensitivity at 35°C. Field of view - horizontally 11.25° and vertically 4°. The detection range for a pedestrian is 300 m.
<i>Fiat and Jaguar</i>	NIR camera placed just above the head of the driver (rear mirror) and light source is over the bar at the front of the car.	Active system NIR. Field of view - horizontally 45°. The detection range for a pedestrian is 150 m.
<i>Autoliv</i>	FIR camera placed at the lower end of the windshield.	Active system NIR. Field of view - horizontally 45°. The detection range for a pedestrian is 150 m.

Display and applications

Initially, it is explored the feasibility that the systems of night vision use a mirror and projector over the dashboard and lower part of the driver's windshield, this unit project real-time thermal images, which appears to float above the hood and below the driver's line of sight. Perhaps, this visualization of the system is good, but to include these devices in the vehicles demand the development of expensive technologies and the users will not pay by them. A more realistic system is currently used in many vehicles; it consists of a liquid crystal display (LCD) embedded in the middle dashboard. The driver check the thermal images and other applications supplied by the vehicle computer.

The current commercial applications, based on infrared images, are limited to display a stream of images, which correspond to events registered in real time by sensors. Although, to develop a system in real time is not simple task, the only operation of image processing is contrast enhancement. Recently, new research lines in night vision develop software that can identify pedestrian or critical situations.

1.2 Multimodal Stereo Head

Computational stereo refers to the problem of determining three-dimensional structure of a scene from two or more images taken from distinct viewpoints. It is a well-known technique to obtain depth information by optical triangulation. Other examples are: stereoscopy, active triangulation, depth from focus, and confocal microscopy.

Stereo algorithms could be classified according to different criteria. A taxonomy for stereo matching is presented by [77]; they propose to categorize them into two groups, which will be explained briefly. *Local methods* attempt to match a pixel with its corresponding one in the other image. These algorithms find similarities between connected pixels through its neighborhood, surrounding pixels provide the information to identify matches. Local methods are sensitive to noise, and ambiguities, such as: occluded regions, regions with uniform texture, repeated patterns, changes of view point or illumination. *Global methods* can be less sensitive to the mentioned problems since high-level descriptors provide additional information for ambiguous regions. These methods formulate the problem of matching in mathematical terms, more than local methods, which allows to introduce restrictions that model surfaces or maps of disparity. For instance: smoothness, continuity, among others. Nowadays, it is a still open research topic to find the best conditions, restrictions, or primitives to decrease the percentage of bad matching pixels. Some methods use heuristic rules, or functional to do it. Their main advantage is that scattered maps of disparity can be completed. This is performed by techniques such as: dynamic programming, intrinsic curve, graph cuts, nonlinear diffusion, belief propagation, deform model, and any other optimization or search procedure².

The existing algorithms also are categorized into different groups, for instance, depending on the number of input images: *multiple images* or *single image*. In the first case, the images could be taken either by multiple sensors with different view points or by a single moving camera (or moving the scene, and holding the sensor fixed). Another classification could be obtained according to number of used sensors: *monocular*, *bifocal*, *trifocal*, and *multi-ocular*. The figure 1.2 shows a generic binocular system with nonverged geometry³.

The fundamental basis for stereo is the fact that every point in three-dimensional space is projected to a unique location in the images (see figure 1.2). Therefore, if it is possible to correspond the projections of a scene point in the images (I_L and I_R), then it is certain that its spatial location on a world coordinate system O will be recovered.

Assuming that: P_L and P'_R are the projections of the 3D point P on the left and right images, and O_L and O_R are the optical center of cameras, on which two reference coordinate systems are centered (see figure 1.3). If also, a pinhole model for the cameras are supposed, and that the image plane arrays are made up of a perfect rectangular grid aligned. Then, the line segment $\overline{C_L C_R}$ is parallel to the x coordinate

²It refers to choosing the best element from some set of available alternatives.

³Camera principal axes are parallel.

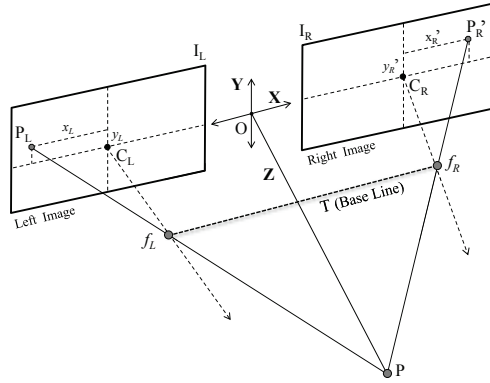


Figure 1.2: A stereo camera setup.

axis of both cameras. Under this particular configuration the point P is defined by the intersecting ray from the optical centers O_L and O_R through their respective images P : P_L and P'_R .

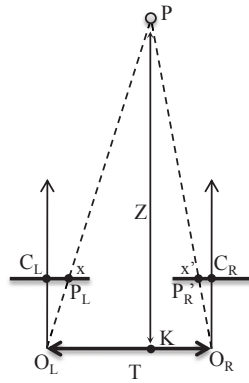


Figure 1.3: The geometry of nonverged stereo.

The depth Z is defined by a relationship of similarity between the triangles $\Delta O_L C_L P_L$ with $\Delta P K O_L$, and $\Delta O_R C_R P'_R$ with $\Delta P K O_R$.

$$\frac{\overline{C_L P_L}}{\overline{C_L O_L}} = \frac{\overline{O_L K}}{\overline{K P}} \quad \text{By similar triangles- } \Delta O_L C_L P_L \text{ with } \Delta P K O_L, \quad (1.3)$$

$$\frac{\overline{C_R P'_R}}{\overline{C_R O_R}} = \frac{\overline{O_R K}}{\overline{K P}} \quad \text{By similar triangles- } \Delta O_R C_R P'_R \text{ with } \Delta P K O_R, \quad (1.4)$$

$$\therefore \overline{K P} = \overline{C_L O_L} \times \frac{\overline{O_L K} + \overline{O_R K}}{\overline{C_L P_L} + \overline{C_R P'_R}} \quad \text{from 1.3 and 1.4} \quad (1.5)$$

$$\text{Or } Z = f \frac{T}{d}, \quad (1.6)$$

where d is the *disparity* or displacement of a projected point in one image with respect to the other; in the *nonverged* geometry depicted in figure 1.3 it is the difference between the x coordinates: $d = x - x'$ (the last one is valid when the pixels x and x' are indexes of a matrix). The *baseline* is defined as the line segment joining the optical centers O_R and O_L .

1.3 Multimodal stereo



Figure 1.4: multimodal stereo head.

This section presents in detail the multimodal stereo head together with the proposed algorithm for computing sparse 3D maps. Figure 1.4 shows an illustration of the multimodal platform. The different challenges of the tackled problem can be appreciated in this illustration, from the image acquisition and depth map estimation to the evaluation of the performance of the algorithm. The different stages of the proposed multispectral stereo are presented in detail below.

1.3.1 Multimodal stereo head

In the current work, a multimodal stereo head with an LWIR camera (PathFindIR from Flir⁴) and a color camera is built. The color camera, by convenience, corresponds to the left camera of a commercial stereo vision system (Bumblebee, from Point

⁴www.flir.com

Grey⁵). The Bumblebee stereo head is used for validating the results and consists of two cameras Sony ICX084 with Bayer pattern CCD sensors, and 6 mm focal length lenses. It is a pre-calibrated system that does not require in-field calibration. In summary, two stereo systems coexist (see Fig. 1.4). The left camera coordinate system of Bumblebee is used as a reference system for both stereo heads. In this way, a kind of ground truth for the depth of each pair of images (infrared and color) is obtained from the Bumblebee stereo head.

The LWIR camera, which will be referred just as LWIR, detects radiations in the range 8 – 14 μm (long-wavelength infrared), whereas the color camera, referred to as VS, responds to wavelengths from about 390 to 750 nm (visible spectrum).

1.3.2 Calibration and rectification

The multimodal stereo head has been calibrated using Bouguet’s toolbox [8]. The main challenge in this stage is to make visible the calibration pattern in both cameras. In order to do this, a special metallic checkerboard has been made using a thin aluminium metallized paper. Black squares over this surface are generated by means of a laser printer, being able to detect them from both VS and LWIR cameras. Figure 1.5(b,d) shows a pair of calibration images (LWIR and color). Despite of using a metallic calibration pattern, the junctions of black and white squares are not correctly detected due to thermal diffusion. Hence, calibration points are extracted using a saddle point detector, instead of a classical corner detector. In our particular case the use of saddle points results in a more stable detection; it is due to the fact that thermal variation between black and white squares are not enough to generate step edges, and the structure of junctions looks more like saddle points than corners [56]. Figure 1.6(a,b) shows three illustrations of junctions obtained with the saddle point detector; note that even though the contrast of these infrared images is different the junctions are correctly detected. Figure 1.6(c,d) depicts local structure indicated by the red windows in Fig. 1.6(a,b); the green points are saddle points while red ones are corners; straight lines show diagonal directions where their intersection corresponds to the most likely position of junctions. As can be seen in these plots, the green points are nearer to the intersections than the corresponding red ones.

Three independent calibration processes under different temperature were performed to study the robustness of intrinsic parameters of LWIR camera when the saddle point detector is used; as a result, the obtained intrinsic parameters were stable beside the changes in temperature. Notice that the LWIR images in Fig. 1.6(a,b) correspond to one image of those calibration sequences.

Once the LWIR and VS cameras have been calibrated, their intrinsic and extrinsic parameters are known, being possible, not only the image rectification, but also to calculate the disparity map of the scene. The image rectification was done, using the method proposed in [29], with an accuracy improvement due to the inclusion of the radial and tangential distortion coefficients into their camera model. An example of

⁵www.ptgrey.com

rectified images is shown in Fig. 1.5(b,d).

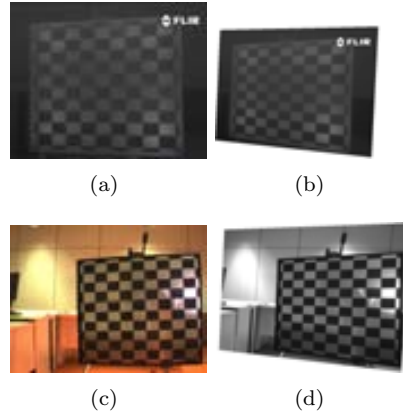


Figure 1.5: (a) Infrared image of the checkerboard pattern. (b) Infrared rectified image. (c) Original color image. (d) Rectified image.

1.4 Evaluation Dataset

1.4.1 Multispectral datasets

Furthermore, a well detailed multispectral dataset together with its corresponding ground truth is proposed. All this material (i.e., multispectral stereo dataset and ground truth images) is available through our website⁶ for an automatic evaluation and comparisons of multispectral stereo algorithms.

a dataset with VS and LWIR images, together with their corresponding disparity maps and 3D models, is publicly available for evaluating different approaches. Up to our knowledge there is not such a kind of dataset in the research community to be used as a test bed.

A multispectral dataset has been generated for evaluating the different stages of the proposed algorithms. It contains multispectral images, ground truth disparity maps and ground truth depth maps. All this information was obtained as indicated below.

The dataset consists of four kinds of images, which are classified by their context and predominant geometry: (i) roads; (ii) facades; (iii) smooth surfaces; and (iv) OSU Color-Thermal dataset. The first three groups were acquired with the proposed stereo head and contain outdoor scenarios with one or multiple planes and smooth surfaces. The latter subset contains perfectly aligned LWIR and color images (i.e., without disparity). It was obtained from [19] and is publicly available⁷. These images

⁶<http://www.cvc.uab.es/adas/datasets/cvc-multimodal-stereo>

⁷<http://www.cse.ohio-state.edu/otcbvs-bench/>

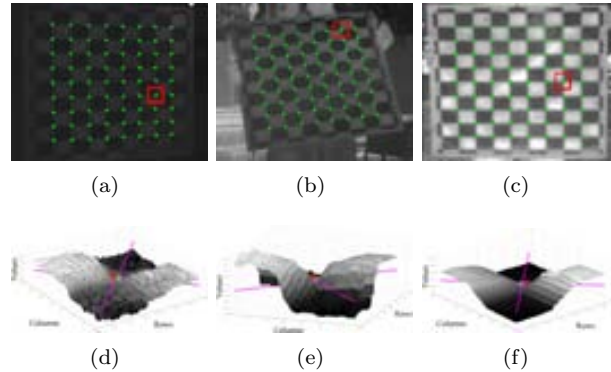


Figure 1.6: Saddle points extracted from infrared images of the checkerboard pattern at different temperatures.

are particularly interesting, since they are aligned the ground truth of disparity maps can be approximated by assuming a registration accuracy of about ± 2 pixels. Figure 1.8 shows an illustration of the whole dataset.

The multispectral stereo images in the dataset have been enriched with ground truth disparity maps and ground truth depth maps semi-automatically generated. These ground truth were obtained by fitting planes to the 3D data points obtained from the Bumblebee stereo head. It works as follows. Firstly, a color image from the left Bumblebee camera is manually segmented into a set of planar regions (see Fig. 1.7(a)). Planar regions are easily identified since are the predominant surfaces in the considered outdoor scenarios. Then, every region is independently fitted with a plane using their corresponding 3D data points, by orthogonal regression using principal components analysis. Figure 1.7(c) shows an illustration of the synthetic 3D representation containing different planes. Additionally, during this semi-automatic ground truth generation process, labels for occluded, valid and unavailable pixels are obtained (see Fig. 1.7(b)). These labels are needed for the evaluation methodology.

Once the 3D planes for a given image are obtained, since they are referred to the VS camera, the corresponding data points are projected to the infrared camera. Thus, a ground truth disparity map is obtained. The fourth column of Fig. 1.8 shows some of these disparity maps and a sparse 3D representation.

In the case of smooth surfaces (e.g., third row in Fig. 1.8) no planes are fitted, and depth information provided by Bumblebee is used as a reference. Bumblebee software offers a trade off between density and accuracy of data points. Hence, in order to have a good representation, its parameters have been tuned so that 3D models are dense enough and contain few noisy data. Those models should not be considered as ground truth, strictly speaking, however we use them as a baseline for qualitative comparisons.

The multispectral stereo head consists of a pair of cameras separated by a baseline of 12 cm and a non verged geometry. This configuration is obtained by adjusting the

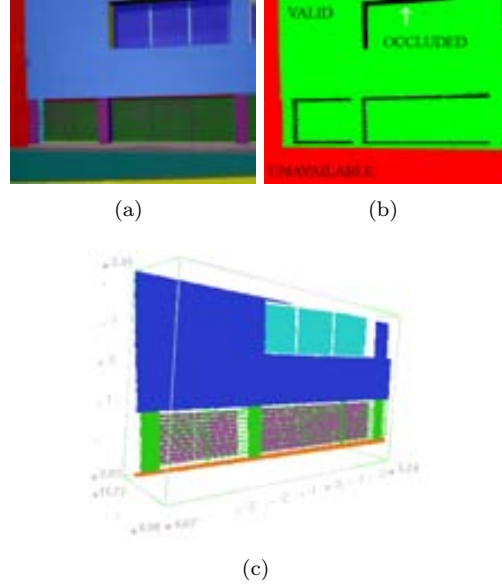


Figure 1.7: (a) Facade image from the proposed dataset overlapped with a mask from the segmentation. (b) Mask of regions with occluded and no depth information. (c) Synthetic 3D representation generated from the visible stereo pair used as ground truth for evaluating the multispectral depth map.

pose of the cameras till their z coordinate axes are parallel, and perpendicular to the baseline. Hence, the images provided by the multispectral stereo head are pre-aligned, ensuring their right rectification. Thermal infrared images are obtained with a *Long-Wavelength InfraRed* camera (PathFindIR from Flir⁸) while color ones with a standard Sony ICX084 camera, which has a focal length of 6 *mm*.

Multispectral stereo camera calibration is considerably more complex than the classical VS/VS, because the LWIR sensor measures heat variations. Therefore, a calibration pattern ideally should have two different temperatures for generating contrast images. In practice, this is not feasible. Furthermore, the effect of thermal diffusion between the calibration pattern and air causes both smooth step edges and distorted corners in infrared images, which are not perceived at a glance. In order to avoid these problems we calibrate the multispectral head in an outdoor scenario using a metallic checkerboard. In this way, sun rays are reflected in white rectangles and absorbed in the black ones, this procedure enhances the contrast of image and helps the detection of calibration points. Although the problem of blurred calibration points is partially solved by the lighting reflection/absorption technique, a saddle point detector is considered instead of a classical corner detector to obtain more robust results.

As mentioned above, the cameras have been aligned before starting the calibration process. This action ensures that the needed projective transformations for their

⁸www.flir.com












Dataset Name	Image			Pairs of Images
	VS	LWIR	Disparity & Depth	
Roads				2
Facades				2
Smooth surfaces				19
OSU Color-Thermal Dataset [19]			—	2

Figure 1.8: Illustration of the four subsets of images contained in the proposed multispectral dataset.

rectification are smooth (the image planes' position are approximately coplanar). Once each camera has been calibrated, and its intrinsic parameters are known, the next step is to estimate the geometry of multispectral stereo rig. Since the current work is focused on the generation of dense disparity maps, it is only necessary to estimate the epipolar geometry (fundamental matrix \mathbf{F}). Then, with this matrix, the next step is to rectify the multispectral images.

The image rectification is a critical issue since the proposed algorithm assumes that all epipolar lines in the multispectral images are horizontally aligned. Despite the accuracy with which \mathbf{F} was estimated, it is essential to use a rectification method that takes into account the large dissimilarity of intrinsic parameters of the cameras. In the current work the method proposed in [66] has been used. This reduces the loss and creation of pixels due to projective transformations during the rectification process (resampling effect), while preserving the aspect of image content.

The disparity maps are provided by a VS/VIS stereo vision system: Point Grey Bumblebee (PGB). Note that, for the sake of simplicity, the multispectral stereo head was presented as an independent system, however it uses one of the cameras that belongs to PGB—the right one. In other words, the camera referred to as VS in the current section corresponds to the right camera of PGB. This stereo rig setup was selected because it is efficient in terms of hardware and software.

Chapter 2

Similarity Window-based Matching Cost Functions

The aim of this chapter is to present a cost function that measures the similarity between two block of data extracted from different modalities; particularly VS and LWIR images. This function combines mutual information with a shape descriptor based on gradient. These metrics are computed for every level of a scale-space representation, and then, they are propagated through that representation following a coarse-to-fine strategy. The main contributions of this chapter are: (i) a complete description of the proposed multimodal cost function, and its relationship with information theory; (ii) a quantitative evaluation of the combined used of gradient and mutual information, and the benefit of their propagation between consecutive levels; and (iii) to establish the principles toward a stereo multimodal cost function.

2.1 Introduction

THE coexistence of infrared cameras with other sensors has opened new perspectives for the development of multimodal systems. One of the challenges is to find the best way to fuse all this information in a useful representation. In the current chapter the problem of corresponding two blocks of data belonging to different spectral bands is considered (multimodal matching problem). These data are provided by two cameras that are capable of measuring emissions in visible and thermal infrared spectrum. Since the cameras are mounted adjacent to each other, offering a global view of objects in the scene, the occlusions are negligible.

The literature on multimodal matching can be broadly divided into *entropy-based methods* and *feature-based methods*. Through this chapter a hybrid approach is proposed. It exploits mutual information and gradient information in scale-space representations.

Mutual information is a concept derived from information theory, and measures the amount of information that one random variable contains about another. It is a powerful concept in situations where no prior relationships between the data are known. The previous property makes mutual information the ideal tool to address problems involving signals without an apparent relationship [33].

Viola *et al.* [94] intuitively introduces mutual information as a measure of alignment between images and 3D models; then, it was formalized in [65], only for images. The importance of these early contributions were to identify the main properties of mutual information in the field of multimodal image processing, and its usability. Just few years later, Engal proposed in [22], a cost function for VS/VS stereo based on mutual information. He assumed that two data blocks are corresponding if the amount of shared information is maximum, that is to say, it is possible to find the correspondence between a window (template) and set of candidate windows (searching space) by maximizing mutual information. Although, its performance is not better in comparison to other less complex cost functions [74, 86], recently; it has been shown that is robust to radiometric differences [42].

Mutual information has been also largely used for medical image registration. In this field, Pluim *et al.* [69] propose to combine mutual and gradient information, showing an improvement with respect to classical mutual information based formulations [65, 85]. Approaches that combine multiresolution schemes with mutual information have also been proposed for local matching of medical imaging. An advantage of these methods is the information suppression, which allow to analyze the structure of the images with different level of details. Thus, the information of the current level can be enriched by using the prior knowledge collected from previous levels in the hierarchy [56, 70]. A strategy similar to the one mentioned above is presented in [27], being restricted to propagation of the joint probability obtained from two patches. In the current chapter not only mutual information (MI) but also gradient information (GI) are propagated through two different scale-space representations. The first one is based on a scale-space stack while the second is a pyramidal representation.

In summary, this chapter presents a quantitative evaluation of performance once MI and GI are considered; Furthermore, a comparison of the discriminative power of a scale-space representation based on stacks and pyramids and the advantages of propagating mutual and gradient information (MI and GI). The proposed approach is evaluated with a large number of experiments; up to our knowledge previous works were, on the one hand, specially devoted to the registration problem; and one the other hand, they were validated on few samples. The rest of this chapter is organized as follows. Section 2.3 presents the theoretical principles of proposed multimodal cost function as well as their relationship with information theory (entropy and mutual information). Experimental results and comparisons are given in Sect. 2.4. Finally, conclusions and final remarks are drawn in Sect. 2.5.

2.2 Matching Cost Volume

The matching cost volume is a three-dimensional array that stores the cost of corresponding two square windows, obtained from a pair of multimodal images. This volume is obtained following a local window based approach, which consists in computing a cost for each displacement of a sliding window, while a second window is kept fixed on a point in the reference image. The cost volume is referred to as $C(\mathbf{p}, d)$, where $\mathbf{p} = (x, y)$ is the point on reference image (I_{VS}) and d is the disparity or the displacement of the sliding window measured in pixel. The point \mathbf{p} corresponds to the center of the squared window, of size wz , placed on $I_{VS}(x, y)$ whereas d represents the location of the sliding window in I_{LWIR} . Specifically, the latter is a window with the same size than the previous one but centered on $I_{LWIR}(x + d, y)$. Notice that the sliding window location can be parametrized by coordinates of the reference window and d since multimodal images are rectified. Finally, the searching space is defined as an interval $[d_{min}, d_{max}]$ that contains all possible values of d . Figure 2.1 shows how a cost $c(x, y, d)$ is indexed in $C(\mathbf{p}, d)$ together with the windows (i.e., reference and sliding windows) used for its calculation.

Based on that representation of the costs, the LWIR/VS correspondence problem could be stated as an optimization problem. In other words, let $C(\mathbf{p}, d)$ be a cost volume. Then, the LWIR/VS *correspondence problem* is equivalent to optimize d for a given \mathbf{p} in a space of candidate solutions C , when a set of constraints J are applied. The sought solution (D) is given by:

$$D = \underset{d}{\operatorname{argmax}}(C(\mathbf{p}, d)), \quad (2.1)$$

subject to: J constraints.

At this point, $C(\mathbf{p}, d)$ could be seen as an *objective function*, *energy function* or *cost function*, which is maximized or minimized, according to the problem. But even more important, all assumptions made by the matching algorithm are translated into an objective function. Furthermore, the correspondence problem could be solved as a constrained optimization on a discrete or continuous domain, depending of the properties of C . In computational stereo, this means D with sub-pixel accuracy.

An advantage of the formulation presented in Eq. (2.1) is that it integrates into a unique representation most of the commonly used assumptions in (VS/VS) stereo [86]. Moreover, sparse and dense solutions are computed in a similar way, only handling the cost volume. Another interesting property of this generic view of LWIR/VS correspondence problem is the processes integration, that is, integration of different tasks into one optimization step. For instance, feature selection and correspondence [64].

Figure 2.1: Multimodal Matching Cost Volume.

2.3 LWIR/VS Matching Cost Functions

The definition of a cost function able to find the good correspondences between information provided by the VS and LWIR cameras is a challenging task due to their poor correlation [68, 78]. In spite of that, recent works on computational stereo [42] have shown that mutual information is a nonparametric cost function able to address nonlinear correlated signals. However, we have found undesirable behaviors when it is used as a cost function in the multispectral stereo problem. Mainly, due to the images are compared only through its information content, ignoring shape information.

This section presents the key concepts involved in the definition of proposed multimodal cost functions. Initially, they are presented as a still equation, those values generating a matching cost volume. Then, their adaptation to the LWIR/VS correspondence problem is individually introduced. They are: (i) mutual information that is presented in [22]; (ii) gradient information; and (iii) their combination (i.e. mutual and gradient information) [69]. Finally, a cost function based on mutual and gradient information in a multi-resolution context is proposed [2].

The discriminative power of *MI* and *GI* is improved through a scale-space representation. Note that a similar scheme is proposed in [27], but propagating joint probabilities ($p(a_i, b_i)$). In contrast, the proposed cost function directly spread the *MI* and *GI* between adjacent levels, allowing changes in the sizes of the bins that represent the sources of information. This supposes a great advantage because at each level an optimum alphabet can be used, which is unsuitable in a scheme such as the one proposed by [27].

Algo para introducir la formula

$$C(\mathbf{p}, d) = [\lambda_0, \dots, \lambda_t]^T \cdot (C_{MI}(\mathbf{p}, d) \cdot \text{diag}(C_{GI}(\mathbf{p}, d))) \quad (2.2)$$

where λ is the confidence of current *MI* or *GI*.

Our approach starts by computing *MI* and *GI* at all scales.

$$C_{MI}(\mathbf{p}, d) = [MI(\nabla_0^0(I(\mathbf{p}, d))), \dots, MI(\nabla_0^t(I(\mathbf{p}, d)))] \quad (2.3)$$

$$C_{GI}(\mathbf{p}, d) = [GI(\nabla_1^t(I(\mathbf{p}, d))), \dots, GI(\nabla_1^t(I(\mathbf{p}, d)))] \quad (2.4)$$

Previous works have shown that the gradient information by itself is not enough to find right correspondence between LWIR and VS images [91, 69], since gradient orientations in the range of $[0, \pi]$ are useful (e.g. half range). Therefore, *MI* helps to *GI* to overcome its loss of descriptiveness. Although there are different ways to combine them, their product is selected, because has a noise cancellation effect, thus low costs are obtained, when LWIR textureless and VS textured regions are put in correspondence (and vice-verse).

2.3.1 Mutual Information

Mutual information has shown to be a valid cost function in several multimodal problems. For instance, medical imaging registration [65, 69], LWIR/VS video registration [52], medical imaging matching [70], and LWIR/VS matching [53]. The results obtained by these studies provide enough evidences to justify its uses in such a kind of problems. Therefore, we formalize the multimodal correspondence problem in context of a more general theory, that is, information theory. For this purpose, first of all, it is necessary to define a probability space (Ω, \mathcal{B}, P) . Formally speaking, this space is defined by Ω that denotes a sample space, \mathcal{B} a space of events defined

as σ -field \mathcal{B} of subsets of Ω , and a probability measure P that assigns a real number $P(F)$ to every member F of the σ -field \mathcal{B} [33]. In our case, I_{LWIR} and I_{VS} are considered as random variables that map thermal infrared or visible measurements (symbols) into finites alphabets A . Thus, $I_{LWIR} : \Omega \rightarrow A_{LWIR}$ and $I_{VS} : \Omega \rightarrow A_{VS}$.

Given that the LWIR/VS images are handled as two information sources, and these produce a succession of symbols in a random manner, the probability space (Ω, \mathcal{B}, P) is a mathematical generalization of the interaction between: (i) symbols that could be intensity or thermal infrared measurements; (ii) the space of all possible output symbols, grouped per blocks; (iii) events, which are sequences of symbols that can be drawn by sources of information; and (iv) a probability measure assigned to every event.

By definition, mutual information (MI) measures the amount of information that one random variable contains about another [16, 33]. It is a useful concept where no prior relationship between the data is known. This is estimated in a local way, for two square windows centered on \mathbf{p} and $(x+d, y)$, and size $wz \times wz$ pixels. Thus, MI is defined as follows:

$$MI(\mathbf{p}, d) = \sum_{a_i, b_j} p(a_i, b_j) \log \frac{p(a_i, b_j)}{p(a_i) p(b_j)}, \quad (2.5)$$

where a_i and b_j are discretized pixel values that are within windows ($a_i \in A_{VS}$ and $b_j \in A_{LWIR}$); $p(a_i, b_j)$ represents their joint probability mass function; $p(a_i)$ and $p(b_j)$ are their respective marginal probability mass functions. The alphabets A_{VS} and A_{LWIR} are built by normalizing each window independently (range $[0, 1]$) and then quantizing them into Q levels. The joint probability mass function $p(a_i, b_j)$ is a 2-dimensional matrix, whose values correspond to the probability that a pair (a_i, b_j) occurs. The marginal probabilities are determined by summing along each dimension of the previous matrix.

An additional explanation of why mutual information is a valid LWIR/VS cost function arises when their boundary conditions are analyzed. Formally, mutual information is bounded to range $[0, \min(MI(\mathbf{p}, \mathbf{p}), MI(d, d))]$ [22], so that, the minimum value occurs when a_i and b_i are completely independent, and it is maximum when these symbols are either identical or they are affected by a transformation T such that generate an equivalent joint probability matrix. The latter boundary condition deserves special attention because this justifies the performance of mutual information as a cost function. Notice that MI is maximized by more than one a_i and b_i combination, therefore, exist a set of T transformation that perform a one-to-one mapping such that $MI(\mathbf{p}, T(\mathbf{p})) = MI(\mathbf{p}, \mathbf{p})$ or $MI(d, T(d)) = MI(d, d)$. This invariance gives a great advantage to MI over other similarity functions, which looking for identical pattern as is common in VS/VS stereo. This enables to MI measure similarity in more situations, particularly when the underlying probability of symbols a_i and b_i is nonlinear.

In order to discuss the capability of mutual information as a LWIR/VS cost function let us consider a simple scenario, like the one the depicted in Fig. 2.2, where a

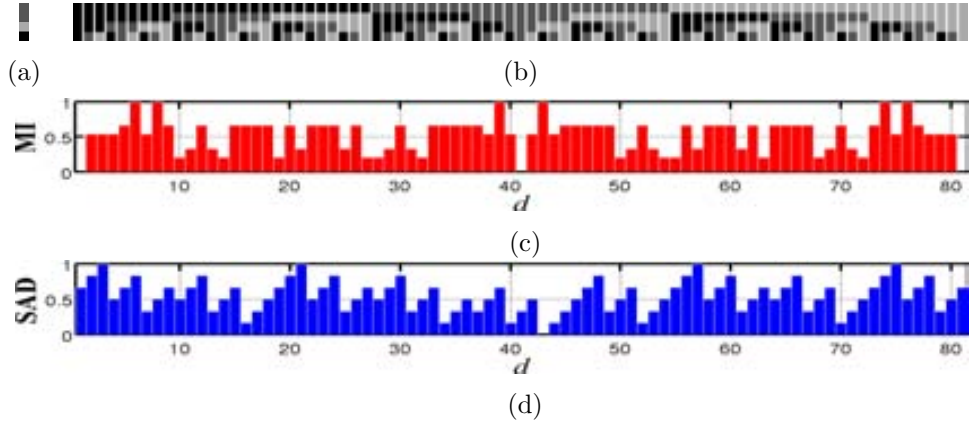


Figure 2.2: Toy example of MI and SAD : (a) patch P ; (b) searching space; (c) mutual information S ; and (d) sum of absolute differences.

patch denoted by P is searched in an image. In this toy example, a random block of 4×1 pixels is discretized into 3 levels, and then is compared with other of the same size extracted from the searching space. This patch is build from an alphabet of 3 symbols, $A_P = \{a_0, a_1, a_2\}$, as can be seen in Fig. 2.2(a). In contrast, the searching space S contains all possible combinations of A_P (i.e. $3^4 = 81$) as is shown in Fig. 2.2(b). In order to find the right match two cost functions are employed. They are (i) mutual information (MI) and (ii) Sum of Absolute Differences (SAD). Their resulting costs are plotted in Fig. 2.2(c) and Fig. 2.2(d), respectively. Potential right match are all those 4×1 blocks at searching space that optimize some of the cost functions, in case of MI they are $\text{argmax}_d(MI(P, S(d)))$, while in the SAD is $\text{argmin}_d(SAD(P, S(d)))$.

It can be observed that SAD function (Fig. 2.2(d)) is minimized just in one position ($d = 43$), actually when a patch identical to P is found in S . On the contrary, MI is maximized several times, in total 6 times, positions $d = \{6, 8, 39, 43, 74, 76\}$. This is due to the fact that the patches on these locations produce equivalent joint probability distributions, and in consequence, equal MI costs. Notice that both MI and SAD costs are normalized to range $[0, 1]$. This simple example illustrates, on the one hand, the main advantage of MI over SAD , and other similarity functions that reward identical pattern, such as NCC , SSD , and PC which is precisely that an identical patch or pattern not necessarily is the right one, because thermal variations have not a direct relationship with intensity variations. So, mutual information is able to find linear and nonlinear correlations between patches, taking into account the whole dependence structure of the variables. On the other hand, in the context of mutual information is clear that a balance between Q and wz is mandatory for accurate solutions.

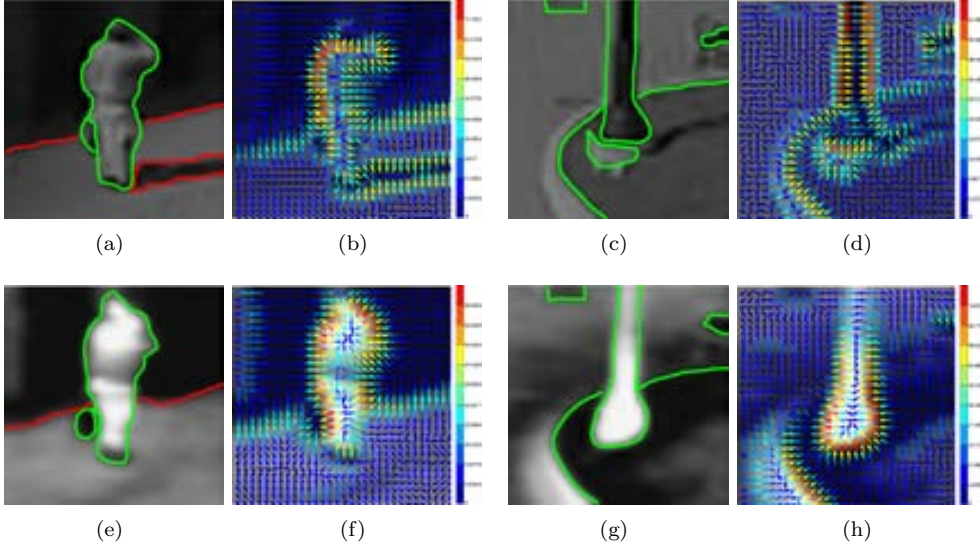


Figure 2.3: Gradient vectors: (a-d) VS patches; and (e-h) LWIR patches.

2.3.2 Gradient Information

The current section explains how the gradient information is incorporated to the proposed LWIR/VS cost function. We adopt the formulation presented in [69], and show that this is not only valid for three-dimensional medical image registration, as was initially proposed, but also is effective for LWIR/VS correspondence problems. Figure 2.3 shows eight image patches and their corresponding gradient vectors. These patches are divided into two groups according to their modalities. So, Fig. 2.3(a-d) depict intensity changes in visible band, while Fig. 2.3(e-h) show thermal infrared variations. Every column depicts a patch from the same scene but from different spectral band, Since these patches are registered (column wise, for example Fig. 2.3(a) and Fig. 2.3(e)) there exist a point-to-point correspondence between pixels and gradients. Every gradient on these figures is represented according to its norm. Thus, strong gradients are red while softs are blue (the gradient color coding is the vertical bar in left side of each patch). Note that in these figures the green and red gradients seem somehow to be related. In fact, experiments conducted by Pluim *et al.* [69, 70] shown a high correlation between middle and strong gradients for medical imaging. In this way, they can assume that image locations with a strong gradient denote transitions, and these transitions have high information content. We come to the same conclusions, however it is necessary to clarify what type of transitions or edges complies with this assumption.

In general, an edge could be classified into three categories [31]: (i) shadow-geometry edges, (ii) highlight edges, and (iii) material edges. However, the wide gap that exists between thermal infrared and visible bands prevents to treat them as a *simultaneous phenomena*, particularly those caused by external agents, such as

illumination. A detectable phenomena in I_{VS} and I_{LWIR} images is denominated as simultaneous *iff* its band-wide to overlap the VS and LWIR bands, otherwise it would be a *typical phenomenon* of a spectral band. Figure 2.3(a) and Fig. 2.3(e) show two shadow edges, one produced by a building (facade), and another by a walking pedestrian. The former appears in both modalities and their gradient are correlated. Whereas, the latter only is registered by the VS camera. This type of shadow edges, which are produced by dynamic bodies are not taken into account by the proposed cost function. Because they are phenomenons that can be perceived only by VS sensor. Furthermore, a moving shadow no alter the heat of a covered surface (by blocking part of the sunlight). The temperature changes in outdoor are typically a slow process, and our LWIR sensor have not the sensitivity for measuring these changes.

The highlight edges are avoided, because often are perceived in different ways by every multispectral sensor. Commonly, highlight and shading in VS spectrum are explained through dichromatic reflection model [79], however this model is unsuitable for the LWIR band, because ignores relevant interactions between body, environment and waves.

Material edges as shown in Fig. 2.3 are strong correlation. For instance: object boundaries, pedestrians, sidewalk, and lamppost.

It is important to note that thermal infrared and intensity variations are not necessarily equals (nor in orientation neither in magnitude). However, since both images depict the same scene, corresponding gradient vectors could appear in both modalities and their phase difference be near to 0 or π (phase or counter-phase). Therefore, these vectors could be used to unveil possible matchings. Let \mathbf{x} and \mathbf{x}' be two corresponding points that belong to I_{VS} and I_{LWIR} , respectively.

Since the images are rectified, not only the search for correspondences is simplified to one dimension, but also the objects in the scene appear with a similar aspect (see Fig.). This is an important fact because the contours and edges are regions with a high LWIR/VS correlation value. Therefore, they have a high probability of being correctly matched [68].

The gradient information is obtained as follows:

$$GI(\mathbf{p}, d) = \sum_{\mathbf{x}, \mathbf{x}'} \underbrace{w(\theta(\mathbf{x}, \mathbf{x}'))}_{\text{Orientation}} \underbrace{\min(|\mathbf{x}|, |\mathbf{x}'|)}_{\text{Norm}}, \quad (2.6)$$

where: $\nabla_1(\cdot)$ is the gradient vector field of I_1 ; \mathbf{x} is a coordinate refereed to this vector field (same for $\nabla_2(\cdot)$, where $\mathbf{x}' \in I_2$); $|\cdot|$ is the norm; $\theta(\mathbf{x}, \mathbf{x}')$ is the angle between them; and $w(\theta)$ is a function that penalizes gradient orientation out of phase or counter phase: $w(\theta) = (\cos(2\theta) + 1)/2$. The gradient information is computed similarly to MI on two windows centered on $\nabla(I_{VS}(\mathbf{p}))$ and $\nabla(I_{LWIR}(\mathbf{q}))$, thus the cost volume $C_{GI}(\mathbf{p}, d)$ is obtained by sliding them through the searching space defined by each \mathbf{p} on the reference image.

where θ is the phase difference of two gradient vectors in the location \mathbf{x} and \mathbf{x}' . It is

defined as follows:

$$\theta(\mathbf{x}, \mathbf{x}') = \arccos\left(\frac{\mathbf{x} \cdot \mathbf{x}'}{|\mathbf{x}| |\mathbf{x}'|}\right). \quad (2.7)$$

The Eq. (2.7) is weighted by a function $w : \theta \rightarrow [0, 1]$ that penalizes those gradient vectors that are not in phase or counter-phase:

$$w(\theta) = \frac{\cos(2\theta) + 1}{2}. \quad (2.8)$$

2.3.3 Mutual and Gradient Information

2.3.4 Scale-Space Context

This section presents the basic notions about scale-space, which are used to build two data structures. Firstly, a scale-space stack representation is presented. Then, a pyramidal representation, which is faster than the previous one, is described.

Stack Representation

The scale-space representation $L : \mathbb{R}^N \times \mathbb{R}^+ \rightarrow \mathbb{R}$ for an arbitrary dimension N is obtained by convolving an image with a Gaussian derivative kernel of order n . Note that the zero scale is also included and corresponds to the given image. Following the notation presented in [62]:

$$L_n(\mathbf{x}; t) = g_n(\mathbf{x}; t) * I_k, \quad (2.9)$$

where $\mathbf{x} = (x_1, \dots, x_N)^T \in \mathbb{R}^N$, $t \in \mathbb{R}^+$ is the current scale level, I_k is a given image, and $g_n(\mathbf{x}; t)$ is the Gaussian derivative kernel of order n . If $n = 0$ the Gaussian function is obtained, otherwise its corresponding derivative kernel. In this chapter, only gradient information is required, hence L_0 and L_1 are computed for I_{VS} and I_{LWIR} . It means a stack of Gaussian blurred images and their corresponding first order derivative images.

Notice that in the case of a stack all the images in the stack have the same size. Thus, $MI(\nabla_0^{t-1}(I_k(\mathbf{p}, d)))$ and $MI(\nabla_0^t(I_k(\mathbf{p}, d)))$ have an interscale correspondence.

Pyramidal Representation

Another way to generate a scale-space representation is by means of a pyramidal hierarchy, which is similar to the method described above. It consists in adding a new stage after the Gaussian filtering, which apply a downscale algorithm, sampling the output image at a constant rate. In this work, we have explored the use of an half

octave Gaussian pyramid of zero and first order [17]. This representation has been chosen due to the reduction factor; hence it assure an optimal propagation of mutual information. Figure 2.4 shows two pyramidal representations of three levels; (a) and (b) correspond to an intensity image, while (c) and (d) to an infrared image. Note that in the two coarser levels the image features are still available, in spite of their small sizes. See [62, 17, 56] for a detailed description on pyramidal representations.

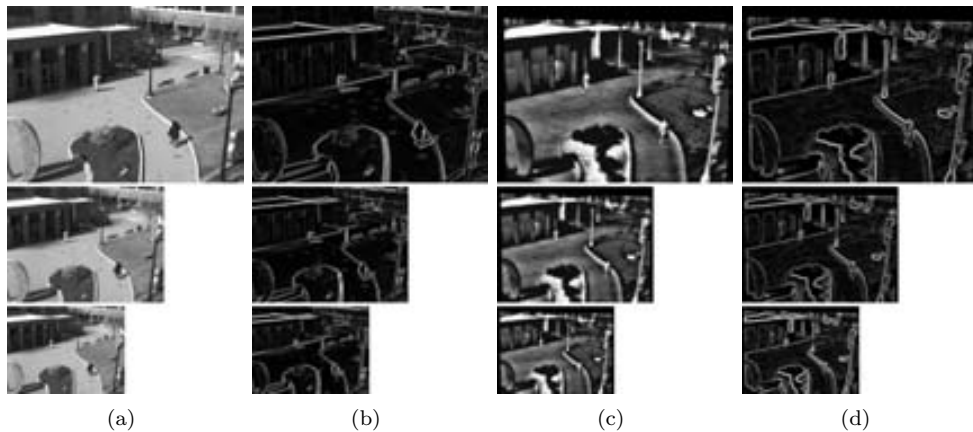


Figure 2.4: Three level pyramidal representations: (a) $L_0(\mathbf{x}; t)$ of I_{VS} ; (b) $L_1(\mathbf{x}; t)$ of I_{VS} ; (c) $L_0(\mathbf{x}; t)$ of I_{LWIR} and (d) $L_1(\mathbf{x}; t)$ of I_{LWIR} .

In the case of a pyramidal representation the level $t - 1$ contains a smaller image than the one in the current level t (downsampling). Therefore, two situations must be considered: (i) if \mathbf{x} is not present in the previous level, only its value at the current level is considered (MI or GI) and the term λ in Eq. (2.2) is set to 1; and (ii) if \mathbf{x} is present in the previous level, then a cubic spline interpolation is used to compute its I_{prior} , since we are using rectified images only ancestors on the epipolar line are considered (one dimensional interpolation problem); thus I_{prior} is obtained from its neighborhood at $(t - 1)$.

2.4 Experiments

In order to evaluate the proposed approach, small parts of an thermal infrared or color images are cropped and used as templates—61600 patterns in total were extracted from OTCBVS Benchmark Dataset [19]. MI and GI are cost functions computed between the template and all possible windows on the corresponding searching space; they are obtained without disparity restrictions. The correct match is located at point d where the cost function reaches the maximum value, as is indicated as follows:

$$D = \underset{d}{\operatorname{argmax}} (C(\mathbf{p}, d)). \quad (2.10)$$

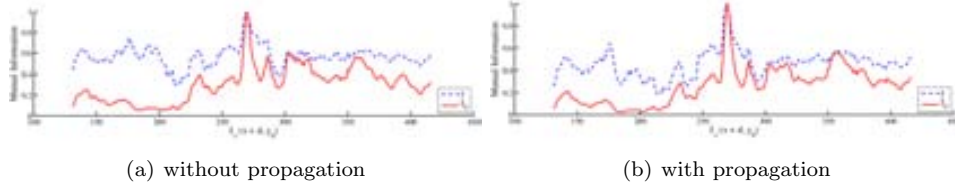


Figure 2.5: (a) Mutual information (MI) and mutual with gradient information (MGI) as formulated in [69]. (b) Proposed propagation of mutual information MI and mutual with gradient information MGI (MI : dashed line; MGI : solid line).

The matching cost of a template and a candidate is obtained by computing MI Eq. (2.5) and GI Eq. (2.4). Once the cost over the whole searching space is computed the three largest local maximum values are extracted (only three values were selected just for the sake of presentation simplicity). These values are used to quantify the results, which are depicted in Table 2.1 and 2.2. Since color and infrared images in [19] are registered, the correct matches are known before hand. Then, it is possible to determine the correct one among the three local maximum selected above (a tolerance range of 2 pixels for d is used). Tables 2.1 and Table 2.4 present the percentages of correct matching that corresponds to first, second or third position. If a *winner – takes – all* scheme was used, then the number of correct matches will be just the first column of these tables. The proposed approaches, both using a scale-space stack and a pyramidal representation, have been compared with the results obtained when MI and MGI are not propagated through the different levels of the stack/pyramid (Table 2.1 and 2.1 part (a)).

Figure 2.5(b) shows the results for the same example introduced above but when MI and GI are propagated. Note that both approaches (with/without propagation) find the correct match but by using propagation the relative values between local maximums are increased, making easier to identify the correct one.

Table 2.1: First three maximums by using a scale-space stack

	1st.		2nd.		3rd.	
	MI	GI	MI	GI	MI	GI
<i>Without propagation</i>						
$t = 2$	37.19	50.03	12.52	11.55	7.08	5.83
$t = 1$	17.30	27.58	9.61	9.95	7.10	5.11
$t = 0$	4.15	12.54	3.62	7.44	3.37	5.92
<i>With propagation</i>						
$t = 2$	37.19	50.03	12.51	12.52	7.08	5.83
$t = 1$	31.69	49.61	12.57	12.57	7.71	5.96
$t = 0$	15.24	43.97	9.19	11.99	7.00	5.74

Table 2.2: First three maximums by using a pyramidal representation

	1st.		2nd.		3rd.	
	<i>MI</i>	<i>GI</i>	<i>MI</i>	<i>GI</i>	<i>MI</i>	<i>GI</i>
<i>Without propagation</i>						
$t = 2$	31.29	54.04	15.86	1386	10.82	8.11
$t = 1$	14.14	29.28	9.66	13.38	7.91	8.95
$t = 0$	4.15	12.54	3.62	7.44	3.37	5.92
<i>With propagation</i>						
$t = 2$	31.29	54.04	15.86	1386	10.82	8.11
$t = 1$	17.67	39.55	15.31	15.53	7.10	8.36
$t = 0$	9.23	27.38	6.41	9.87	5.15	6.52

Upper levels of scale-space stacks were obtained by convolving the images with a Gaussian kernel of order $n = \{0, 1\}$ and $\sigma = \{1, 2\}$, as shown in Table 2.1. The experiments were conducted following the next setup, in both the stack and the pyramid cases. The window size decreases with the scale. It started with a size of 32×32 and finishes with 8×8 (level 0); the propagation also follows this direction. The parameter λ controls the degree of propagation between consecutive levels. Experiments have shown that $\lambda = 0.5$ maximizes the scores. The quantization parameter Q is constant ($Q = 30$).

MI and *GI* showed a behavior proportional to the size of template (\bar{I}_l). If it is increased then the estimation of *MI* will be better, due to large number of observations. Nevertheless, big windows are not desirable for stereo matching. Therefore, our propagation scheme is a good choice because it improves the results whereas small windows (8×8 pixels) are used. The improvement obtained with the scale-space stack reaches about 3.5 times at the last level, while in the pyramidal representation it is about 2.2 times due to the downsampling.

The representations only have three levels in order to compare both results. The results of pyramid using propagation are better than without it. However, these results cannot be compared to the ones obtained with the stack, except at level 0, due to compression of images (see Table 2.2). Notice that, each level contains less information and the image is smaller; hence, the estimation of *MI* is weak. The used mutual information estimator (Eq. (2.5)) and the way to ensemble the alphabets, establish a dependency between the estimation (*MI* value) and the number of members in the sample (template size), which affects the performance of propagation in this representation.

Results presented in Table 2.1 and Table 2.2 show the improvements reached when gradient information is used with mutual information, instead of mutual information alone. On average, *MGI* improves the result from *MI* about 3 times.

2.5 Conclusions

This chapter presents a scheme for combining mutual information with gradient information together with an evaluation of two scale-space representations. Experimental results show the improvements in the discriminative power as well as the viability of the proposed approach. Future work will study a mutual information estimator robust to downsampling.

Chapter 3

Multimodal Sparse Stereo

This chapter presents an imaging system for computing sparse depth maps from multispectral images. A special stereo head consisting of an infrared and a color camera defines the proposed multimodal acquisition system. The cameras are rigidly attached so that their image planes are parallel. Details about the calibration and image rectification procedure are provided. Sparse disparity maps are obtained by the combined use of mutual information enriched with gradient information. The proposed approach is evaluated using a Receiver Operating Characteristics curve. Furthermore, a multispectral dataset, color and infrared images, together with their corresponding ground truth disparity maps, is generated and used as a test bed. Experimental results in real outdoor scenarios are provided showing its viability and that the proposed approach is not restricted to a specific domain.

3.1 Introduction

THE coexistence of visible (VS) and thermal infrared (LWIR) cameras has opened new perspectives for the development of multimodal systems. In general, visible and infrared cameras are used as *complementary* sensors in applications such as video surveillance (e.g. [58, 11]) and driver assistance systems (e.g. [45]). Visible cameras provide information at diurnal scenarios while infrared cameras are used as night vision sensors. More recently, Near-InfraRed (NIR) and visible images have been successfully used in tasks, such as image registration [24], scene category recognition [10], and removing shadows [73]. These works assume that NIR and VS images could be registered, and that exist a correspondence pixel-to-pixel between red, green, blue and NIR channels, which allows to develop *cooperative* frameworks that overcome state-of-the-art algorithms operating in VS spectrum.

All the approaches mentioned above involve registration and fusion steps, resulting in an image that even though contains several channels of information lies in the

2D space. The current work goes beyond classical registration and fusion schemes by formulating the following question: “*is it possible to obtain 3D information from a multispectral stereo rig?*”. It is clear that if the objective is to obtain depth maps close to state-of-the-art, classical binocular stereo systems (VS/VS) are more appropriated. Therefore, the motivation of current work is to show that the generation of 3D information from images belonging to different spectral bands is possible. The proposed multispectral stereo rig is built with two cameras, which are rigidly mounted and oriented in the same direction. These cameras work at different spectral bands, while one measure radiation in the visible band the other one registers infrared radiation. From now on, this system will be referred to as *multimodal stereo head*, which is able to provide a couple of multispectral images.

The role of cameras in the proposed multimodal stereo system is not only restricted to work in a *complementary* way (as it is traditionally) but also in a *cooperative* fashion, being able to extract 3D information. This challenge represents a step forward for the 3D multimodal community, and results obtained from this research by sure can benefit applications in the driver assistance or video surveillance domains, where the detection of an object of interest can be enriched with an estimation of its aspect or distance from the cameras.

The performance of a stereo vision algorithm is directly related to its capacity to find good correspondences (*matching*) between pairs of images, this task relies on the similarity function used to match features. In the multimodal case, similarity functions, such as *SAD* (sum of absolute differences), *NCC* (normalized cross correlation), *SSD* (sum of squared differences) or Census transform cannot be used since a linear correlation between the data cannot be assumed [38]. In the current work a non linear similarity function, that establish the relationship between multimodal images is adopted. In other words, it is able to associate information content between LWIR and VS images.

Multimodal matching has been widely studied in registration and fusion problems, specially in medical imaging (e.g., [4, 76, 83]). However, there are few research related with the correspondence problem when thermal infrared and color images are considered. Hence, it is not clear how to exploit visible and infrared imaging in a cooperative framework to obtain 3D information.

Most of the stereo heads presented in the literature, and other commercially available, are built from cameras that have the same specifications (i.e., sensor and focal length). This choice constrains the problem and facilitates the reuse of software and published methods. However, the case tackled in the current work is far more complex since heterogeneous sensors are used, besides the intrinsic problems due to multimodality. So, the alignment of two views coming from cameras with different sensors and intrinsic parameters should be taken into account, which is more difficult than a classical VS/VS stereo heads.

The use of multimodal stereo heads (LWIR/VS) has attracted interest of researchers in different computer vision fields, for examples: human detection [37], video surveillance [54], and 3D mapping of surface temperature [97, 71]. Recently, [53] presents a comparison of two stereo systems, one working in the visible spectrum

(composed of two color cameras) and the other in the infrared spectrum (using two LWIR cameras). Since the study was devoted to pedestrian detection, the authors conclude that both, color and infrared based stereo, have a similar performance for such a kind of applications. However, in order to have a more compact system they propose a multimodal trifocal framework defined by two color cameras and a LWIR camera. In this framework, infrared information is not used for stereoscopy but just for mapping LWIR information over the 3D points computed from the VS/VS stereo head. This allows to develop robust approaches for video surveillance applications (e.g., [54]).

On the contrary to the previous approaches, a multimodal stereo head constructed with just two cameras: an infrared and a color one is presented in [52]. This minimal configuration is adopted in the current work since it is the most compact architecture in terms of hardware and software. Critical issues such as camera synchronization, control signaling, bandwidth, image processing, among other have a minimal impact in the overall performance, and can be easily treated by an acquisition system such as the one presented in [63]. In Krotosky et al. [52] this compact multimodal stereo head (LWIR/VS) is used for matching regions that contain human body silhouettes. Since their contribution is aimed at person tracking some assumptions are applied, for example a foreground segmentation for disclosing possible human shapes, which are corresponded by maximizing mutual information [94]. Although, these assumptions are valid, they restrict the scope of applications.

A more general solution should be envisaged, allowing such a kind of multimodal stereo head to be used in different applications. In other words, the matching should not be constrained to regions containing human body silhouettes. The current chapter has two main contributions. Firstly, a robust approach that allows to compute sparse depth maps from a multimodal stereo head is proposed. Since it is not restricted to a specific application it can be used in any scenario. Finally, the second contribution is the adaptation to the multimodal case of a recently presented methodology for comparing and evaluating stereo matching algorithms. This evaluation method has been proposed for classical stereo heads where both cameras work in the same spectral band [51]. It is based on Receiver Operating Characteristics (ROC) curves that capture both error and sparsity.

Although the proposed approach is motivated for recovering 3D information, optionally it could help to solve other multimodal problems. Due to the fact that most existing multimodal systems are affected by the same problem. That is, statistical independence between the different modalities, which makes difficult its correlation. Our approach offers a non-heuristic based solution, which is a novel feature with respect to state of the art. The current section presents an approach that reveals the information shared by the modalities, and from these correspondences find the match between blobs or image regions. The latter is relevant for multimodal applications such as moving target detection, medical imaging, video fusion, among other.

The paper is organized as follows. Section 3.2 presents details about the generation of matching cost volume. Section 3.3 describes the optimization steps and constraints applied for computing sparse depth maps. Section 3.4 introduces the evalua-

tion methodology. Experimental results with different scenes are presented in Sect. 3.5, together with the technique used for setting the parameters of the algorithm. Conclusions and final remarks are detailed in Sect. 3.6.

3.2 Matching Cost Volume Computation

A crucial aspect of the multimodal stereo algorithms is to find good matching, despite of the poor correlation between LWIR and VS images [78, 68]. This problem is tackled by using of C_{MGI} , a cost function that combines mutual and gradient information in scale-space context. Although, this similarity function is defined for multimodal template matching (Chapt. 2), the current chapter shows that also can be used as a cost function in a multimodal stereo system. Additionally, its accuracy is improved through a Parzen window estimator, as will be described below.

The join probability p_{a_i, b_j} from Eq. (2.5) is estimated in two steps. Firstly, it is constructed a two dimensional histogram of discretized pixels a_i and b_j . Every entry is obtained as follow:

$$p(a_i, b_j) = \frac{1}{wz^2} \sum_{\mathbf{x}, \mathbf{x}'} T[(a_i, b_j) = (\mathbf{x}, \mathbf{x}')], \quad (3.1)$$

$T[\bullet]$ is a conditional function, which takes the value of 1 if the argument between bracket is true, and 0 otherwise. It is appropriate to recall that both \mathbf{x} and \mathbf{x}' represent a pair corresponding points in local window coordinates, and that wz is the size of those windows. Once all entries are computed, the joint probability is approximated by a Parzen estimator as is presented in [48]. It assumes a Gaussian distribution g with standard deviation σ_g on every entry of the previously obtained histogram ($p(a_i, b_j)$). Thus:

$$p(a_i, b_j) = p(a_i, b_j) * g(a_i, b_j, \sigma_g). \quad (3.2)$$

The other probabilities from Eq. (2.5), are determined by summing along each dimension of the previous joint probability (see [94] for more details about the probability estimation).

Similarly to cost volume presented in Chapt. 2, a multimodal cost volume $C_{MGI}(\mathbf{p}, d)$ is computed $\forall \mathbf{p} \in I_{VS}$. This representation comes from Eq. (2.3) and Eq. (2.4), when a smooth joint probability is used for computing MI .

In contrast to template matching algorithm presented in Sect. 2.4, two stereo images are not similar. Therefore, the calibration and rectification have a decisive impact, because not only the search for correspondences is restricted to one dimension, but also the contours and edges of the objects contained in the scene have a similar aspect. This fact increases the probability of coincidence on contours and boundaries as is shown in [68].

3.3 Disparity and Depth Computation

The process of disparity selection consists of two steps. Initially, the disparities with higher cost values are selected as correct with a classical winner take all criterion. In these cases, a correct match is determined by the positions d (image coordinate) where the cost function reaches the maximum value: $\arg \max_d (C_{MGI}(\mathbf{p}, d))$. The disparity map obtained after this first step contains several wrong entries due to regions with low texture or no information. Note that the multispectral stereo matching case is more complicated than traditional (VS/VS) ones. The latter is due to the fact that, for instance, an object in the scene could appear textured in the visible spectrum, while it could have the same temperature all over its surface, therefore appear as a constant region in the infrared image, and vice versa.

As mentioned above, it is hard to select the correct d between several candidates with similar scores. Therefore, a second step to reject mismatching candidates is added. It consists in labelling as correct those correspondences with a cost score higher than a given threshold τ_{MGI} . The selection of this threshold is based on error rates (see Section 3.4). Next, these reliable matchings are used for bounding the searching space in their surrounding. As it will be shown in the 3D maps, this helps to discard wrong matchings.

The τ_{MGI} parameter is included into our formulation for picking up only those pixels with large C_{MGI} values. Since the C_{MGI} cost function is reliable in textured regions, and those regions have higher C_{MGI} cost, τ_{MGI} is used as a threshold that split up the cost map into two groups: (i) reliable matches and (ii) unreliable matches. This parameter exploits the correlation between material edges.

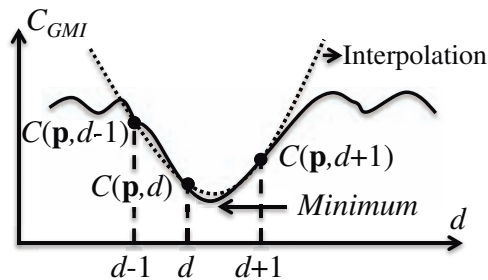


Figure 3.1: Disparity interpolation.

Finally, a quadratic curve is used for a fine estimation of disparity values; this function fits a polynomial to points $\{d - 1, d, d + 1\}$ and its respective cost values (see Fig. 3.1). After computing the disparity of every point in the images (color and thermal infrared), their corresponding 3D positions (X, Y, Z) are obtained using a standard function for triangulation, which is included into the calibration toolbox [8].

3.4 Evaluation Methodology

The current section describes the quality metrics used for evaluating the performance of the proposed multimodal stereo matching algorithm. This metric is inspired by a technique recently presented for classical (VS/VS) semi-dense stereo matching algorithms.

In general, stereo algorithms have been evaluated following the methodology proposed in [74], which has become in a *de facto standard* in the stereo community. It presents two quality measures: (i) RMS (root mean squared) error; and (ii) percentage of bad matching pixels. In both cases, resulting disparity maps are compared to a known ground truth in a dense fashion. However, in uncontrolled scenarios, as outdoors, trying to get ground truth data as presented in [41] or [75] is not feasible, for that reason, we must evaluate our proposed algorithm following a semi-dense methodology.

The method presented in [51] capture both, error and sparsity in a single value, which is suitable for our dataset. So, we extend this framework to the multispectral case. The pairs: *error* and *sparsity* are plotted in a Receiver Operating Characteristics (ROC) curve as a unique value, letting visualize how performance is affected as more disparity values are taken. Remember that every disparity obtained by our method have a cost value associated, which depends on C_{MI} and C_{GI} . Therefore, regions with low information (low entropy) or without texture (gradient) could be rejected considering their cost. During the evaluation process the best τ_{MGI} parameter could be easily identified (see Sect. 3.3).

In the current section, ROC curves have been used for evaluating the performance of the proposed approach independently of parameter settings $\{wz, Q, \sigma_t\}$. The evaluation procedure is briefly detailed below following the original notation.

The statistics about the quality of a semi-dense stereo algorithm should capture both: (i) how often a matching error happens and (ii) how often a matching is not found. These two values define the Error Rate (*ER*) and the Sparsity Rate (*SR*) respectively. In other words, the *ER* represents the percentage of incorrect correspondences:

$$ER = \frac{\text{incorrect_correspondences}}{\text{all_matchable_pixels}}. \quad (3.3)$$

On the other hand, the *SR* is defined as the percentage of all missing correspondences over the set of matchable pixels:

$$SR = \frac{\text{missing_correspondences}}{\text{all_matchable_pixels}}. \quad (3.4)$$

Note that these values are not computed over the whole set of pixels but over those pixels with a match in the ground truth. An illustration of ROC curves, for different scenarios, can be seen in Fig. 3.1 (the meaning of these plots will be explained in

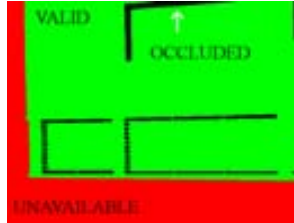


Figure 3.2: Evaluation regions.

Sect. 3.5). These plots have four extrema points: the origin, which represents a dense and perfect matching algorithm; its opposite, where no correct matches are found; the $(0, 1)$ point corresponds to an algorithm that is dense but fully wrong; and finally, the $(1, 0)$ point that corresponds to a disparity map completely empty.

The evaluation by ROC curves compares row by row, a horizontal profile belonging to ground truth disparity map with its corresponding one obtained by the tested algorithm. A correct matching is assumed when the difference with respect to the correct value is smaller than or equal to 1 pixel. Note that only three sets of images (i.e., roads, facades and OSU) are used for the evaluation to avoid the problem of occlusion, which is slightly different in the VS/VS and VS/LWIR stereo rigs. Regarding the *roads* dataset, in all the image pairs there is a single plane hence there are not occluded areas; while in the *facades* dataset occluded areas are removed by generating a synthetic 3D model. Let us remember that OSU dataset was not obtained with our multispectral stereo rig; it is provided by [19] and contains perfectly aligned VS and LWIR images. The *smooth surfaces* dataset is not used during the evaluation since the differences between occluded areas in VS/VS and VS/LWIR stereo rigs could affect the results. Hence, the *smooth surfaces* dataset is just used for a qualitative validation of the proposed approach.

Figure 3.2 shows three kind of regions identified in our dataset: *Occluded*; *Unavailable* (e.g., no textured or too far / close to the multispectral stereo head); and *Valid* regions. A region is valid when depth information is known or is possible to fit a plane with its defining pixels. Therefore, let V be the set of all pixels in ground truth with disparity information available; O be the occluded regions; B be the regions close to an occlusion, by definition, this boundary is 5 pixels of wide; and finally, C be the candidate matches obtained by the evaluated algorithm.

On beginning of this section is introduced the concepts of the two error metrics, ER and SR . Now, they are defined as a function of the following two terms. The operator $T[\bullet]$ that is defined in Eq. (3.2), and a image coordinate \mathbf{p} that lie both on the ground truth, and on the disparity map obtained by the proposed approach. Notice that, ground truth and disparity map are referred to the same coordinate system. Thus, they can be overlapped and their coordinates are equivalents.

Mismatch (M): a correspondence with a disparity value different from the ground truth value larger than one pixel:

$$M(\mathbf{p}) = T[|V(\mathbf{p}) - C(\mathbf{p})| > 1], \quad (3.5)$$

this score considers pixels near to occlusions (B).

False Negative (FN): an unassigned correspondence where a ground truth data is available (i.e., a hole):

$$FN(\mathbf{p}) = T[\mathbf{p} \in V : \mathbf{p} \notin C]. \quad (3.6)$$

False Positive (FP): an assigned correspondence in occluded areas:

$$FP(\mathbf{p}) = T[\mathbf{p} \in C : \mathbf{p} \notin V]. \quad (3.7)$$

The ROC space is defined by the above functions, and from them ER and SR are obtained; remember that they are used as vertical and horizontal axis respectively, in the ROC plots.

$$ER = \frac{1}{|V|} \sum_{\mathbf{p}} (M(\mathbf{p}) + FP(\mathbf{p})), \quad (3.8)$$

where \mathbf{p} only takes the values of valid images coordinates (see Fig. 3.2) and $|V|$ is the number of valid pixels. Finally:

$$SR = \frac{1}{|V|} \sum_{\mathbf{p}} FN(\mathbf{p}). \quad (3.9)$$

In the ROC curves presented in Fig. 3.1, the sparsity rate parameter is varied as follows: the cost values of the candidate matches in C are sorted in descending order. Next, from this list, and by using a decreasing τ_{MGI} threshold, different values of the ROC curve are obtained. For instance, the first plotted element in the ROC curve corresponds to bottom right point, which is the maximum cost value achieved only for a few set of pixels. Then, by decreasing the τ_{MGI} threshold, all the other points that define the ROC curve are obtained. In other words, the more pixels are selected reducing the sparsity rate, the larger the resulting error rate.

3.5 Experiments

This section presents experimental results obtained with different algorithm settings and scenes. The setting of parameters is obtained from two optimization steps. The first one is intended to find the best setting of: (i) window size, (ii) scale and (iii) quantization levels, from the parameter space $P = \{wz \times \sigma_t \times Q\}$. The second optimization step is devoted to find the best confident value $\Lambda = \{\lambda_0, \dots, \lambda_{t-1}\}$ used

for propagating C_{MI} and C_{GI} costs through consecutive levels (2.2). These two steps have been implemented as follows.

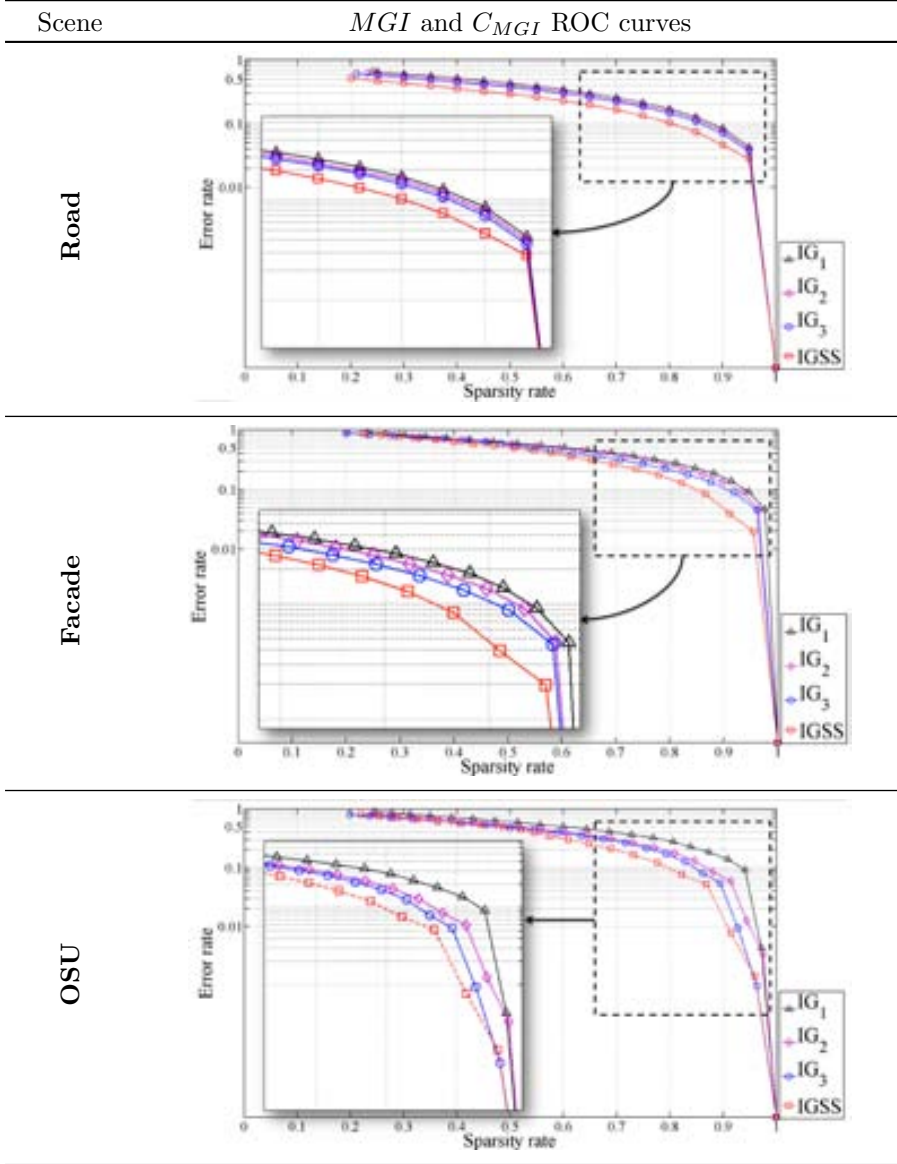
Firstly, an efficiency measure (em) is defined to be used as a quantitative value for comparisons between different settings. Let $em = \int_0^1 ER d_{SR}$ be the area under the error curve (ER) defined for all SR in the interval $[0, 1]$, for a given setting of parameters. The parameter space P is sampled in a limited number of values defining a regular grid. Then, the best setting of parameters corresponds to the node of that grid where em reaches the minimum value. Since in the proposed approach a scale space representation is used, not only the setting with the minimum em value is considered, but the best p_i settings. Note that no prior information about the number of levels in the scale space representation is assumed. Hence, the family of parameter settings, with the lowest error, is obtained. This first optimization step is performed for each subset of the whole dataset. By analyzing the results it is possible to find similarities between the best settings for the images in the evaluation dataset. Thus, it is possible to find relationships between the elements of the parameter space, particularly the relationship between the window size and the quantization level.

Then, the second optimization step finds the best set of $\Lambda = \{\lambda_0, \dots, \lambda_{t-1}\}$ values for merging the C_{MI} and C_{GI} costs corresponding to each of the p_i settings obtained above. Although initially a large family of p_i settings were considered, we experimentally found that three levels were enough to propagate the C_{MI} and C_{GI} costs through the scale space representation. Hence, this second optimization process finds the best $[\lambda_0, \lambda_1, \lambda_2]$ using a similar approach.

The two optimization steps mentioned above are used to find the best combination of parameter settings. Initially, an exhaustive search in parameter space P is performed. The results are used to illustrate the behavior of ER and SR in each subset of dataset. Figure 3.1 shows the three error curves corresponding to: road, facade, and OSU color-thermal. These curves depict the error and sparsity rate when the best settings are used in C_{MI} and C_{GI} costs function (Eq. (2.3) and Eq. (2.4), respectively), together with the improvement achieved by merging them (C_{MGI}). Finally, after finding the best settings for the whole dataset (including confidence parameters $[\lambda_0, \lambda_1, \lambda_2]$) several sparse depth maps of real outdoor scenarios are presented (see right columns in Fig. 3.2 and Fig. 3.3).


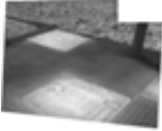
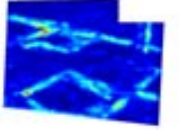


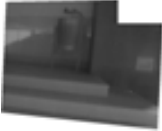
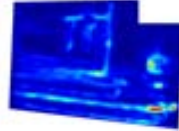



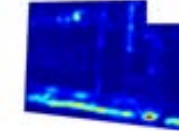


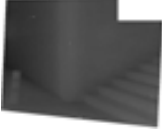
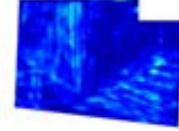


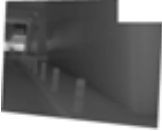
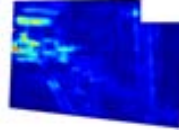


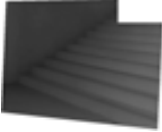
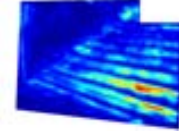

The settings of parameters corresponding to the ROC curves presented in Fig. 3.1 were found with an exhaustive search in the following ranges: $wz = \{7, 19, 31\}$, $\sigma_t = \{0.5, 1, \dots, 6\}$ and $Q = \{8, 16, 32, 64\}$. The best set of parameters and propagation scheme is the following: $MGI_2(p_2) \rightarrow MGI_1(p_1) \rightarrow MGI_0(p_0)$, where $p_2 = \{31, 1.5, 32\}$, $p_1 = \{19, 1, 16\}$ and $p_0 = \{7, 0.5, 8\}$. In our proposal, the windows sizes (wz parameter) decreases from 31 to 7 pixels, which looks like an inverted pyramid. This is to avoid smooth disparity maps, specially on edges, contours and boundaries, since the smaller windows (7×7) contributes in the last stage. On the other hand, we observed that information content decreases with scale, as previously reported in [56], but in our case faster at $\sigma_t = 2$. So, σ_t greater than this value decreases the correct matching score. This is due to the fact that gradient is not enough discriminative (GI in Eq. (2.6)), and the windows tend to have low entropies (MI

Table 3.1: Results obtained at different scales and with different settings ($MGI_2(31, 1.5, 32)$, $MGI_1(19, 1, 16)$ and $MGI_0(7, 0.5, 8)$; as well as their merging, C_{MGI} , with the proposed scale space representation).




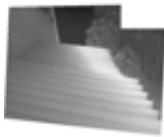
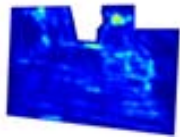



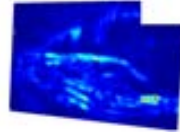



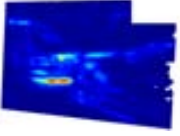



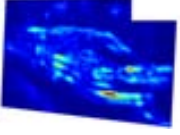



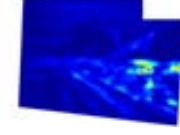



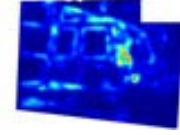

in Eq. (2.5)). Therefore, the propagation of costs should be done with appropriate parameters since a wrong setting could increase the error rate. Regarding λ , the best settings, for the above parameter space $P = \{p_2, p_1, p_0\}$, is $\Lambda = \{0.45, 0.30, 0.25\}$.

Table 3.2: Examples of sparse depth maps from outdoor scenarios (red color in cost map corresponds to high cost values).

VS	LWIR	C_{MGI} cost map	3D results
			
			
			
			
			
			

At this point, we must distinguish between MI coming from the discrete version Eq. (3.2) and its smooth version obtained from Eq. (3.2). The difference lies in how the joint probability $p(a_i, b_j)$ is estimated. When the discrete joint probability is used, MI results in a wave-like curve difficult to minimize. On the contrary, when the smooth $p(a_i, b_j)$ is considered a better behaved function is obtained, which helps

Table 3.3: Examples of sparse depth maps from outdoor scenarios (red color in cost map corresponds to high cost values).

VS	LWIR	C_{MGI} cost map	3D results
			
			
			
			
			
			

us to find stable parameters. Parzen window estimation is done using three different Gaussian kernels (see Eq. (3.2)): $g(a_i, b_j, 3)$ for $wz = 7$; $g(a_i, b_j, 7)$ for $wz = 19$; and $g(a_i, b_j, 9)$ for $wz = 31$.

As a conclusion from the plots presented above we can mention that the best result is obtained when the C_{MGI} is used. Improving the result at each scale, in a coarse to fine scheme. On average a 20% of correct matches, with less than 10% of ER , can be obtained with the proposed approach and by setting the parameters as indicated above. Another conclusion from Fig. 3.1 is that for a given sparsity rate always the best results (lowest error rate) is obtained after merging MGI with the proposed scale space representation C_{MGI} .

Figures 3.2 and Fig. 3.2 show the results obtained with the proposed method. First and second columns correspond to the rectified images, visible and thermal infrared, respectively. Third column presents cost maps obtained after applying disparity selection method explained in Sect. 3.3. Each pixel in this representation corresponds to the maximum C_{MGI} value for a given d that maximize $\arg \max_d (C_{MGI}(\mathbf{p}, d))$. Finally, the fourth column depicts the 3D sparse depth maps obtained from the correct matches. Sparse maps show how the multimodal correspondence between LWIR and VS can provide useful 3D information. Notice that the complexity of images used for validating the proposed approach, is also a challenge for a VS/VS stereo algorithm. However, the obtained results demonstrate the capability of our approach for finding correspondences in a wide range of radiometric differences, such as uncontrolled lighting conditions (sources), vignetting and shadows. Furthermore, the experimental results correspond to outdoor scenes with non-Lambertian surfaces and weakly textured regions. The construction of such a challenging dataset is motivated to push the limits of this novel technique, and provide insights of its application and research trends.

The results shown in the tables must be understood beyond the sparseness of 3D representations, or the accuracy with which the contours are recovered. For example, notice that the vehicles in the LWIR images appear quite poor textured, whereas in the VS images they appear textuless, however our approach can overcome this situation and provides a depth map free of mismatches over those regions (the same for the contrary case). This is consequence of the manner in which mutual and gradient information are combined. Thus, the multiplication of I and G reaches its maximum when a given correspondence is weighted as a correct one by both MI and GI cost functions equation A more dense representation could be obtained by relaxing the τ_{MGI} threshold, but it will be affected by noisy data. Actually, this is a common trade off in stereo vision systems. It should be noticed that 3D representations presented in Fig. 3.2 and Fig. 3.3 provides not only the (X, Y, Z) and color components (r, g, b) , as classical stereo systems, but also the thermal infrared information corresponding to every 3D point.

The cost maps presented in Fig. 3.2 and Fig. 3.3 show that, in general, the cost function introduced in Chapt. 2 can match a pair of window extracted from multimodal stereo image with different accuracy. Our algorithm is designed to identify regions with high information content, such as edges and contours, and from them to

obtain a 3D representation of the scene. Also, it penalizes mismatches in textureless areas, which are not reliable to find correspondences, for instance in image regions such as walls and floor. As can be appreciated on Fig. 3.2 and Fig. 3.3 (third column), higher cost values are concentrated on the edges, since in those regions a consensus between MI and GI is reached. Furthermore, it is possible to perceive the structure of the scene from these cost maps, which confirms the importance of discontinuities for relieving the ill-posedness of multimodal stereo. The strategy of cost propagation across a scale space representation enriches the C_{MGI} , allowing to identify the correct disparity of a candidate set (Sect. 3.3).

As a result, we can affirm that although the current section is focused on recovering 3D information, we have confirmed that the C_{MGI} cost function overcomes mutual information and gradient-based approaches in multimodal template matching problems. This conclusion is supported by reviewing previous work [2], which uses a similar cost function. Since both evaluations (the current and previous one) use the same database (OSU Color-Thermal dataset [19]), we conclude that C_{MGI} is a valid cost function for searching correspondences in multimodal video sequences. This conclusion could be also extended to the multimodal pedestrian tracking and detection problem. The previous statement is motivated by the fact that the work of Krotosky *et al.* (e.g., [54], [53]) is based only on the use of mutual information as a similarity function for matching pedestrian regions.

Finally, regarding the question formulated in Sect. 3.1: “*is it possible to obtain 3D information from a multispectral stereo rig ?*”, we can say with safety that it is possible and it represents a promising research topic with several open issues.

3.6 Conclusions

This chapter presents a novel multimodal stereo matching algorithm of color and infrared images. The different stages for obtaining sparse depth maps are described. Furthermore, a ROC-based evaluation methodology is proposed for evaluating results from such a kind of multimodal stereo heads. It allows to analyze the behavior over a wide range of different parameter settings. Although the obtained results show a sparse representation, we should have in mind the challenge of finding correspondences in between these two separated spectral bands.

In summary, the main contributions of the current work are: (i) to present a study in an emerging topic as Multimodal Stereo LWIR/VS and achieves a sparse 3D representation from images coming from heterogeneous information sources; (ii) to propose a consistent criteria for making the multimodal correspondence; (iii) to establish a baseline for future comparisons; and (iv) to propose a framework that can be used as a test bed for evaluation purposes in this field.

Next sections will be mainly focused on two aspects: (i) improving the disparity selection process by including Markov Random Fields, which allows to consider prior knowledge of the scene; and (ii) reformulating C_{GMI} as a combination of two individual cost functions, which convert the cost function from a consensus scheme to a

scheme where MI and GI contributes to a final matching score according to a set of assignment weights.

Chapter 4

Piecewise Planar Stereo

This chapter proposes a new framework for extracting dense disparity maps from a multispectral stereo rig. The system is constructed with a thermal infrared and a color camera. It is intended to explore novel multispectral stereo matching approaches that will allow further extraction of semantic information. The proposed framework consists of three stages. Firstly, an initial sparse disparity map is generated by using a cost function based on window matching in a multiresolution context. Then, by looking at the color image, a set of plane hypotheses is defined to describe the surfaces on the scene. Finally, the previous stages are combined by reformulating the disparity computation as a global minimization problem. The chapter has two main contributions. The first contribution combines mutual information with a shape descriptor based on gradient in a multiresolution scheme. The second contribution, which is based on the Manhattan-world assumption, extracts a dense disparity representation using the *graph cut* algorithm. Experimental results in outdoor scenarios are provided showing the validity of the proposed framework.

4.1 Introduction

The development of multimodal systems has been an attractive research topic in the computer vision field during the last decade; mainly because they provide a rich representation of the scene by means of a collection of images taken by different sensors. These systems have grown to become a significant tool for dealing with a wide range of problems, for instance: remote sensing, navigation, surveillance, medical imaging, among others. However, in the 3D information recovery domain the potentiality and capability of such systems are still not clear. In the current chapter, a multimodal stereo matching algorithm for extracting dense disparity maps from thermal infrared and color images is presented. These images are acquired with a Long Wave Infra-Red band camera (LWIR) and a color camera (VS) respectively.

Thermal infrared/visible multimodal 3D representations can be broadly divided into two categories according to the role performed by the LWIR camera. The first category includes systems that combine thermal infrared cameras with well-studied techniques for extracting 3D, such as stereo-vision systems (VS/VS) or structured light. These systems are responsible for providing depth information, which is then enriched with the thermal measurement (e.g., [93] and [97]). Although, a valid multispectral representation of the given scene is achieved, the thermal information is treated as a *complementary* source. That is, only mapping thermal infrared information into the resulting 3D representation. On the contrary, the second category includes those approaches where thermal and visible information is matched for extracting a sparse 3D representation (e.g., [52] and [2]). In other words, the information is used in a *collaborative* framework.

Up to our knowledge dense disparity maps only from the first category have been reported in the literature. However, the increasing number of systems where LWIR and visible cameras coexist leads us to state the following questions: “is it possible to obtain *dense* disparity maps from a multispectral stereo head defined with a camera working in the visible and another in the thermal infrared spectral band?”. From an efficiency point of view we wonder whether these *complementary* sensors could be used in a *cooperative* framework that allows to exploit thermal and visible information for extracting a 3D representation.

The structure of the paper is the following. A review of related work on multispectral stereo algorithms is presented in Sect. 4.2. Then, the steps of the proposed algorithm are presented in Section Sect. 4.3. Technical details of multimodal stereo head used for evaluating the proposed approach are presented in Sect. 4.4, together with details of the generated data set and the obtained experimental results. Conclusions and final remarks are given in Sect. 4.5.

4.2 Background

The extraction of 3D information from multispectral stereo heads (LWIR/VS) has attracted the interest of researchers in different computer vision applications, for examples: human detection [37], video surveillance [54], and 3D mapping of surface temperature [97], [71]. Recently, a comparison of two stereo systems is presented in [53], one working in the visible spectrum (composed of two color cameras) and the other in the infrared spectrum (using two LWIR cameras). Since that study was devoted to pedestrian detection, the authors conclude that both, color and infrared based stereo, have a similar performance for such a kind of applications. However, in order to have a more compact system they propose a multimodal trifocal framework defined by two color cameras and a LWIR camera. In this framework, infrared information is not used for stereoscopy but just for mapping LWIR information over the 3D points computed from the VS/VS stereo head. This allows to develop robust approaches for video surveillance applications (e.g., [54]).

On the contrary to the previous approaches, a multimodal stereo head constructed

with just two cameras, an infrared and a color one is proposed in [53]. In this case the challenge is to match regions that contain human body silhouettes. Since their contribution is aimed at person tracking, some assumptions are applied, for example a foreground segmentation for disclosing possible human shapes, which are corresponded by maximizing mutual information [94]. Although, these assumptions are valid, they restrict the scope of applications to those scenarios containing pedestrians. Furthermore, it should be noted that this approach is able to extract 3D information only on those pixels defining the surface of the pedestrian's body.

A more general solution should be envisaged, allowing such a kind of multispectral stereo head to be used in different applications. In other words, the matching should not be constrained to regions containing some predefined characteristic (e.g., human body silhouettes). Note that formulating the solution in a general framework is mandatory in order to extract dense disparity maps.

Up to our knowledge, none of the previous multispectral stereo algorithms for thermal infrared and visible images are able to obtain dense representations. Although the proposed framework is based on a Manhattan World assumption, which could be seen as a constraint, it should be noted that piecewise planar representations are valid in most of man-made environments [14].

In the current chapter, we are interested in generating dense disparity maps from a minimal multispectral stereo head; a similar problem has been addressed in [55]. This study tests two energy minimization scheme only based on mutual information, as are presented in [39, 48]. Its main conclusion is that energy functions based on mutual information terms are not appropriated for solving the thermal infrared and visible correspondence problem. Thus, minimization methods, such as graph cuts [48] and semi-global matching [39], do not converge to the optimal solution since it is difficult to predict the value of a pixel in other spectral band when its corresponding one is given (the well knowing thermal infrared and visible spectrum correlation problem [68]).

Clearly, some constraints must be imposed in order to overcome the low correlation between thermal infrared and visible bands. In the current chapter, these constraints are collected into the so called: *Manhattan world assumption* [15]. These constraints have been originally used to model urban environments with building or city landscapes, exploiting regularities of contours and boundaries. Although, the search of correspondences between stereo images of these environments share classical common problems such as: lack of textured areas, occlusions, shadows, strong lighting changes, repetitive patterns (e.g., bricks facade), constant colored regions (e.g., painted walls). The use of region-based algorithms that assumes Manhattan environments have shown effectiveness in this kind of scenes (e.g., [28, 67]). Our results confirms that an algorithm similar to these provide a interesting road to LWIR/VS stereo.

Region-based algorithms are not new, they have been proposed for VS/VS stereo matching, and roughly these consist of two steps. Firstly, a pre-initialization step is performed to compute a sparse representation. Then, a global minimization scheme is applied to generate the sought dense solution. These algorithms exploit the structural regularities and sharp edges, which are a common characteristics in Manhattan world

scenes.

We found out that region-based methods based on the manhattan world assumption and those that constraint the scene to a set of plane structures (i.e., [6, 49, 95]) can be adapted to the thermal infrared and visible correspondence problem. The key aspect of this adaptation is the cost function used for measuring the similarity between the images. In the current chapter a multimodal cost function, similarly to the one introduced in Chapt. 1, is used. It includes a significant change in its formulation, which improves the accuracy in presence of multispectral edges and textureless regions.

Once an suitable LWIR/VS cost function is established, it is possible to obtain an initial 3D sparse representation of the scene, where the edges are the predominant characteristic. This representation is used to unveil the structure of the scene and as initialization of the proposed region-based algorithm.

Summarizing, The main contribution of this chapter is to propose a complete framework for dense disparity map estimation assuming a piecewise planar model of the scene. The proposed framework spans from an acquisition module up to the proposed algorithm for obtaining dense disparity maps (Chapters 2 and 3). The contributions are as follow: (i) it propose a multimodal cost function to exploit reliable similarities between multimodal images; and (ii) it proposes a standard Markov Random Field for obtaining dense disparity maps, which represents the scene by means of a piecewise planar model.

4.3 Piecewise Planar Stereo

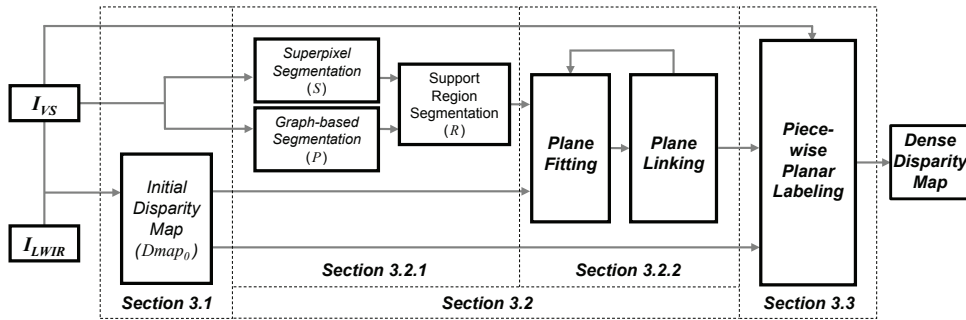


Figure 4.1: Illustration of the algorithm's stages.

The proposed approach consists of three stages. Firstly, it starts by estimating a sparse but accurate disparity map of the scene. Then, in the second stage the initial map is represented by means of a set of planes. Finally, a dense disparity map is obtained by a piecewise planar labeling framework. These stages are detailed next; an illustration is provided in Fig. 4.1.

4.3.1 Initial Disparity Map

The goal of this section is to compute an initial disparity map, which will be fitted by a set of planes. This disparity map is obtained from a matching cost volume following a local window based approach, in which a Winner Take All (WTA) strategy is used for disparities selection. The main challenge of this first stage are both, to get a large number of good correspondences and to have a high accuracy in their locations.

In order to address the matching problem, we propose to extend a cost function based on mutual information by enriching it with gradient information in a scale space representation [2]. A motivation of this proposal is shown in Fig. 4.2(a) and Fig. 4.2(b), LWIR does not match at a pixel level with VS; so classical stereo strategies cannot be directly applied. On the contrary, mutual information, as shown in similar multispectral problems (e.g., [70] and [27]), can be used in this case. Furthermore, in the same figure we can see that edges seem to be a relevant feature present in both modalities; this motivates us to include this kind of information in the proposed solution. Finally, a scale space representation adds robustness and spread local matches from coarser to finer scales increasing the number of final correspondences. In order to tackle the second challenge mentioned above, related with the accuracy of the locations, disparity values are obtained by local quadratic interpolations.

The initial disparity map is obtained by using a matching cost function inspired by the one presented in Chapt. 2. Particularly, we replaced the fusion operator and redefine the system of weights, from scales to information source. Thus, the parameter λ weighted the contribution of C_{MI} and C_{GI} given a scale, instead of its combined contribution (C_{MGI}) scale by scale. This chapter follows nomenclature introduced previously. Thus, it is assumed that I_{VS} and I_{LWIR} are rectified, the searching space constrained to one dimension, the image coordinate $\mathbf{p} = (x, y)$ corresponds to a given pixel in I_{VS} , while the locations $(x + d, y)$ represent to their candidates correspondences in I_{LWIR} , and d stands a disparity value. The cost of corresponding two windows centered on points $I_{VS}(\mathbf{p})$ and $I_{LWIR}(x + d, y)$ is obtained as follows:

$$C(\mathbf{p}, d) = \lambda C_{MI}(\mathbf{p}, d) + (1 - \lambda)C_{GI}(\mathbf{p}, d), \quad (4.1)$$

where C_{MI} and C_{GI} are the cost terms based on the mutual and gradient information given a scale space representation, which will be detailed next; and $d = \{d_{min}, \dots, d_{max}\}$.

By definition in information theory, Mutual Information (MI) measures the information content in common between two random sampled signals, considering them as a collection of symbols that are drawn in a random manner [16]. However, from the point of view of our problem those samples are a pair of windows centered on $I_{VS}(\mathbf{p})$ and $I_{LWIR}(x + d, y)$ respectively, which encode energy measurements in visible and thermal infrared bands. Similarly, we propose to use MI as a cost function that assigns a value depending on its information content; in other words, probabilities of symbols. Formally, $MI(\mathbf{p}, d)$ is defined in terms of individual entropies $h(\cdot)$ and joint entropy $h(\cdot, \cdot)$ as:

$$MI(\mathbf{p}, d) = h(\mathbf{p}) + h(d) - h(\mathbf{p}, d). \quad (4.2)$$

Alternatively, the above equation can be expressed in its continuous form as integrals of the marginal probability distribution functions (PDFs) and joint PDF of pixel values i_{VS} and i_{LWIR} into I_{VS} and I_{LWIR} respectively, then:

$$h(\mathbf{p}) = - \int_0^1 P(i_{VS}) \log(P(i_{VS})) di_{VS}, \quad (4.3)$$

$$h(d) = - \int_0^1 P(i_{LWIR}) \log(P(i_{LWIR})) di_{LWIR}, \quad (4.4)$$

$$h(\mathbf{p}, d) = - \int_0^1 \int_0^1 P(i_{VS}, i_{LWIR}) \log(P(i_{VS}, i_{LWIR})) di_{VS} di_{LWIR}, \quad (4.5)$$

where $P(i_{VS}, i_{LWIR})$ represents the joint PDF, $P(i_{VS})$ and $P(i_{LWIR})$ two marginal PDFs. Kim et al. [48] approximate these PDF and PDFs by a Parzen window density estimation, which is a sum of Gaussian distributions g , with mean μ and covariance ψ (a detailed explanation can be found in [94]). In the current section, a *nonparametric estimator (NP)* [21] is used for computing the joint PDF, instead of using a Parzen estimator. In this way, we avoid dependencies in the selection of parameters: μ and ψ of the Parzen estimator and the parameter needed for binning the windows (Q). Notice that a joint PDF is a two dimensional histogram, where rows and columns represent symbols within windows. In our scope, these symbols come from pixel values of multispectral images; however, since thermal infrared measurements tend to be concentrated in few bins, particularly in outdoor scenes where the temperature remains uniform (thermal equilibrium), the contribution of the estimator used in the current section is significant because it does not require a parameter tuning for binning as Barrera *et al.* [2]. Hence, the joint PDF is obtained as:

$$P(i_{VS}, i_{LWIR}) = NP(\mathbf{p}, d). \quad (4.6)$$

Once $P(i_{VS}, i_{LWIR})$ is obtained, the joint entropy term, $h(\mathbf{p}, d)$ in Eq. (4.5), is computed as follows:

$$h(\mathbf{p}, d) = - \sum_{i_1, i_2} \log(P(i_{VS}, i_{LWIR})) * g_\psi(i_{VS}, i_{LWIR}), \quad (4.7)$$

where g_ψ is a Gaussian kernel needed to approximate the continuous form in Eq. (4.5) from to its equivalent discrete (see [48] and [40] for more details). In practice, we found

that using a small kernel of 3×3 pixels is enough for achieving good approximations, despite of the few samples in windows. Finally, marginal PDFs, corresponding to $P(i_{VS})$ and $P(i_{LWIR})$ in Eq. (4.3) and Eq. (4.4), respectively, are computed in a similar way to the joint probability but along each dimension of $P(i_{VS}, i_{LWIR})$. Thus, $P(i_{VS}) = \sum_{i_{LWIR}} P(i_{VS}, i_{LWIR})$ and $P(i_{LWIR}) = \sum_{i_{VS}} P(i_{VS}, i_{LWIR})$. Then:

$$h(\mathbf{p}) = - \sum_{i_{VS}} \log(P(i_{VS})) * g_{\psi}(i_{VS}), \quad (4.8)$$

$$h(d) = - \sum_{i_{LWIR}} \log(P(i_{LWIR})) * g_{\psi}(i_{LWIR}). \quad (4.9)$$

Note that $P(i_{VS})$ and $P(i_{LWIR})$ are one dimensional vectors, and g_{ψ} also a one-dimensional Gaussian kernel.

Mutual information finds linear and nonlinear correlations between a pair of windows, taking into account the whole dependence structure of variables. However, since local image structures provide rich information that could be also exploited, we introduce a term based on gradient information (*GI*). Thus, this new term is intended to contribute to the matching score in textured regions comparing the orientation of gradient vectors. It is based on the observation that gradient vectors with similar orientations unveil potential matches.

The gradient information presented in Eq. (2.6) also is employed without changes in the current section. This remains being the product of two elements: the first one measures the similarity in the orientation of gradient vectors; while the second one is a factor that weights this similarity value (for more details see Sect. 2.3). Furthermore, the propagation scheme of *MI* and *GI* presented in previous chapter is also incorporated, but they are combined at the lowest scale. In contrast with the initial formulation, which combining them in every scale. Thus, mutual and gradient information cost are defined as follows:

$$C_{MI}(\mathbf{p}, d) = [\alpha_0, \dots, \alpha_t]^T \cdot [MI(\nabla_0^0(\mathbf{p}, d)), \dots, MI(\nabla_0^t(\mathbf{p}, d))], \quad (4.10)$$

$$C_{GI}(\mathbf{p}, d) = [\beta_0, \dots, \beta_t]^T \cdot [GI(\nabla_1^0(\mathbf{p}, d)), \dots, GI(\nabla_1^t(\mathbf{p}, d))], \quad (4.11)$$

where t is an index that refers to a level in the scale-space ($t \in \mathbb{N}$); ∇_0^t and ∇_1^t are scale-space representations given by convolution of a image with a Gaussian kernel of standard deviation (σ), which is progressively increased until obtaining an image stack. Two Gaussian derivative kernels of order 0 and 1 are used to generate blurred and gradient stacks. The α_i and β_i are weights for the linear combination of the results from the different levels of the stack.

The cost volume refereed in Eq. (4.1) is computed from C_{MI} and C_{GI} , Eq. (4.10) and Eq. (4.11), respectively. Section 4.4 presents all values of the parameters above described, and how are they set. Finally, the initial sparse disparity map ($Dmap_0$)

is extracted from C_{MGI} using the strategy of minimization based on WTA criterion with bounded searching space presented in Sect. 3.3. Figure 4.2(c) shows an illustration of the sparse disparity map resulting from this first stage (output) as well as the multimodal input images Figs. 4.2(a,b).

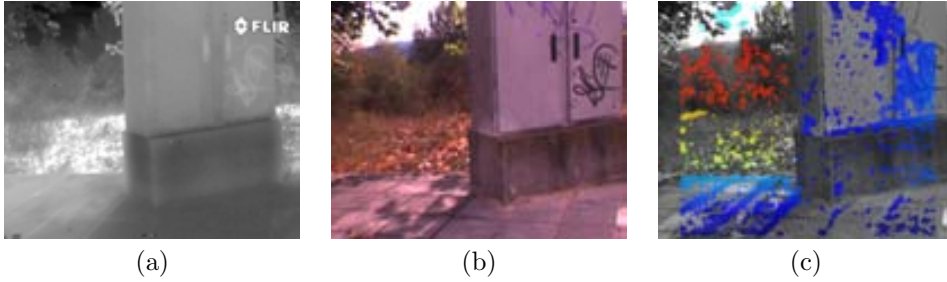


Figure 4.2: Inputs and output images of first stage: (a) rectified infrared image I_{LWIR} ; (b) rectified visible image I_{VS} ; and (c) initial sparse disparity map $Dmap_0$.

4.3.2 Plane Based Hypotheses

In this section the given color image is segmented into a set of regions. Then, each region is represented by a single plane, using the information from the initial sparse disparity map. These planes are used in the final stage as labels for computing the dense disparity map we are looking for.

Support Region Segmentation

This step involves the combination of two segmentation algorithms (i.e., [57] and [23]), which are applied to I_{VS} for obtaining regions that preserve the objects boundaries in the scene. These segmentation algorithms are used in a split-and-merge scheme, in order to unveil potential planar regions. So, the image is decomposed into small regions (superpixels) that later on are connected, following a perceptual criterion. It should be noted that this combination of algorithms is motivated by the application domain (man-made environments).

The segmentation into support regions begins by splinting the given I_{VS} into a large set of small regions, referred to as *superpixels* [57]: $S = \bigcup_i s_i$. Without loss of generality, we assume that disparity values inside a superpixel s_i can be accurately fitted by a plane; this assumption is met as long as I_{VS} is oversegmented. Hence, a trade-off between size of superpixels and fitting error should be found (in the current work 500 regions were used). Large regions have a high probability of covering more than a single planar surface, on the contrary, smaller ones provide few samples to make a proper estimation of the geometry of the selected region. Then, in order to extract perceptually meaningful regions, the segmentation algorithm proposed in [23] is applied to I_{VS} . This results in a set of P partitions of the reference image:

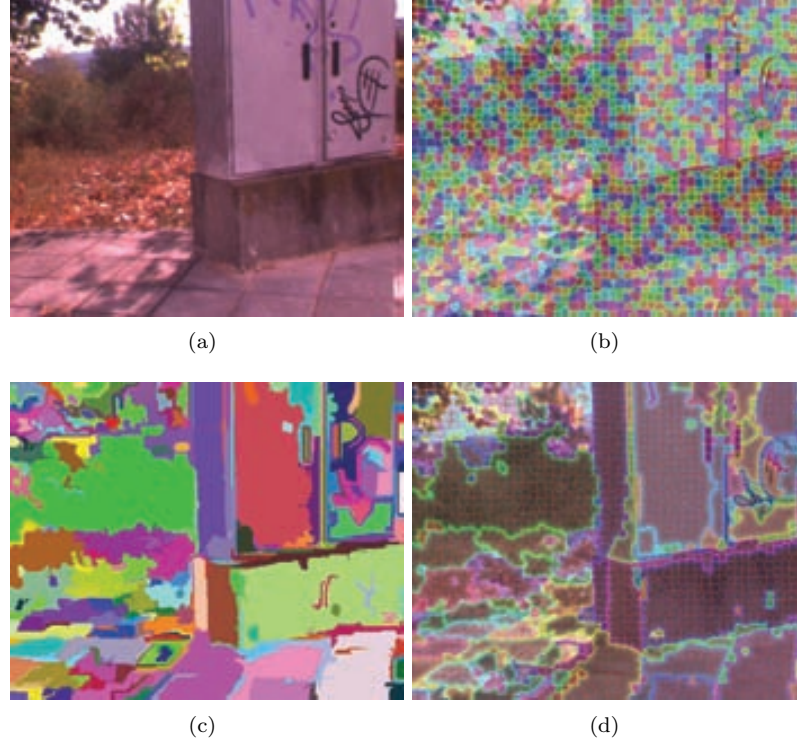


Figure 4.3: Illustration of the support region segmentation: (a) original I_{VS} ; (b) superpixels (S) obtained from [57]; (c) perceptual regions (P) from [23]; and (d) support regions (R) obtained by fusing (b) and (c).

$P = \bigcup_i p_i$. Finally, the results from superpixels (S) belonging to the same perceptual region (P) are connected giving rise to the support regions $R = \bigcup_i r_i$ we were looking for (details on the two segmentation algorithms can be found in [57] and [23] respectively):

$$r_i = \bigcup_{j \in \Omega_i} s_j, \quad \Omega_i = \{j \mid s_j \cap p_i \geq s_j \cap p_k, k \neq i\} \quad (4.12)$$

where Ω_i are the indexes of those superpixels with a maximum overlapping with the given perceptual region p_i . The combined use of those segmentation algorithms (i.e., [57] and [23]) is considered because theoretically guarantee a stable segmentation that preserves the region boundaries and it adapting to local structure of scene. Other algorithms of segmentations or combinations also are valid. For instance, the mean shift algorithm by Comaniciu *et al.* [13] commonly is employed in VS/VS region-based stereo algorithms, such as [7, 95, 6], but it could lead to under-segmentation in the absence of boundary cues, as is reported in [57]. I_{VS} is selected for segmentation because the world coordinate system is set in VS camera and there are a large amount

of algorithms and code available (in contrast to LWIR imaging). Finally, Fig. 4.12 shows an illustration of the results from the two segmentation algorithms, S and P , as well as their fusion R .

Planar Hypothesis Generation

Once the sparse disparity map ($Dmap_0$) has been computed and the color image segmented into r_i regions, a set of hypotheses of planar regions to describe the surfaces in the scene is imposed. So, for every region $r_i \in R$ a RANSAC like algorithm [25] is employed to estimate a pair $\langle \widehat{n}, \bar{x} \rangle$, where \widehat{n} is the normal vector and \bar{x} is the mean value coordinates of the points used for fitting this plane. Note that the planar region estimator operates in the disparity space (x, y, d) , which is different to previous approaches that work on depth maps represented in the Euclidean space (e.g., [30] and [82]).

A RANSAC based plane estimator is chosen since the accuracy of the sought disparity maps depends directly on the confidence of the planar hypotheses. By definition, these methods are capable to find local models from noisy cloud of data; for instance, previous works have demonstrated that this kind of algorithms overcomes least squared based techniques, since they are less sensitive to outliers [92]. It should be mentioned that only those regions r_i that contain three or more valid disparities ($Dmap_0(r_i)$) are considered during this fitting step.

Once RANSAC algorithm has been applied to all the regions in R , a postprocessing stage is performed to merge planar patches defined by similar parameters. This postprocessing is performed to simplify the number of planar hypotheses. Note that the planes have been obtained in a local way, then the number of planar hypotheses could be as large as the number of regions in R . Hence, the goal of this postprocessing stage is to reduce the number of planar hypotheses $\Pi = \{\pi_1, \pi_2, \dots, \pi_n\}$ up to a minimum value so that the structure of the scene is still preserved. The plane linking stage is based on a distance ($dist_\Pi$) computed from two planar patches, which was initially proposed in [88]. It is defined as follow:

$$dist_\Pi(\pi_i, \pi_j) = l(\pi_i, \pi_j) + l(\pi_j, \pi_i), \quad (4.13)$$

$$l(\pi_i, \pi_j) = \frac{(\bar{x}_j - \bar{x}_i) \cdot \widehat{n}_j}{\widehat{n}_i \cdot \widehat{n}_j}. \quad (4.14)$$

The Eq. (4.14) corresponds to the length of the segment defined by \bar{x}_j and the intersection of \widehat{n}_j , passing through \bar{x}_j , with π_i . In order to make it clear, a 2D representation of the segment lengths used for computing Eq. (4.13) is given in Fig. 4.4.

The previous planes distance Eq. (4.13) is used as a similarity function for merging a pair of planar patches. Hence, two planar patches are fused into a single one if ($dist_\Pi(\pi_i, \pi_j) \leq \tau_{LINK}$). Once all possible combinations have been evaluated (only connected neighbor regions are considered) a new relabelled R is obtained and the RANSAC algorithm is called again until convergence is reached.

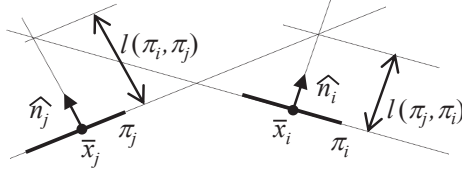


Figure 4.4: 2D illustration of segment lengths used to compute $dist_{\Pi}(\pi_i, \pi_j)$.

Finally, once there are not more planes to be joined, a noisy planar hypothesis removal is performed. It is based on detecting the predominant planar orientations, using a PCA over all normal vectors (\hat{n}). This filtering stage tends to remove planes with an orientation \hat{n}_i far away from the principal directions. This results in a compact set of planar hypotheses $\Pi = \{\pi_1, \pi_2, \dots, \pi_c\}$; it is expected that the number of hypotheses has been reduced: $c \ll n$. Figure 4.5(b) shows the planar hypotheses obtained after merging planar patches with similar parameters and filtering the noisy ones. The original set contains 179 hypotheses (see Fig. 4.5(a)), while the one presented in Fig. 4.5(b) is defined by only 14 hypotheses. They were obtained after four iterations of the plane linking stage.

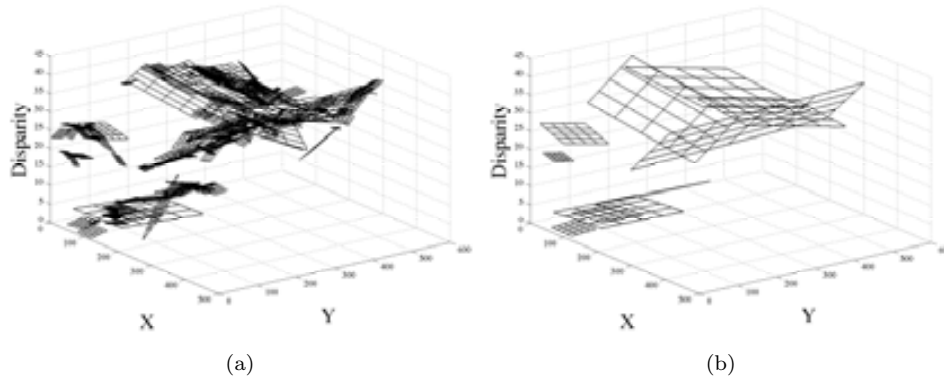


Figure 4.5: Planar hypotheses simplification: (a) original set of planar hypotheses from the segmentation presented in Fig. 4.3 (179 planes) and (b) planar hypotheses resulting after four iterations of the postprocessing stage (14 planes).

4.3.3 Piecewise Planar Labeling

The set of planar hypotheses obtained above are now converted into labels for reformulating the disparity computation as a global minimization problem. It allows to take into account contextual constraints in order to achieve a dense disparity representation from multispectral information. The global minimization problem is based on the local correlation indicators computed in previous sections (i.e., mutual and gra-

dient information boosted by the scale-space representation). In this section, former indicators that were extracted at a level of pixels, are now interpreted as projections of planar surfaces. This helps to constrain the searching space to a few candidates, while spatial coherence of disparity values is hold. Notice that an extra planar hypothesis denoted as π_∞ that represents all those regions out of the stereo range is added to Π (e.g., sky or distant surfaces).

The Markov Random Field theory provides a framework to relate local correlation indicators together with contextual constraints. These two elements are used to define a energy function. Then, this function is minimized through the classical graph cuts [9]. It works by defining a regular grid (four-connected) where every node represents a pixel in the image. Hence, a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, where \mathcal{V} represents the vertices and \mathcal{E} represents the edges of the graph, is obtained. Then, the graph cut algorithm searches in \mathcal{G} for the best set of labels (f) that minimizes the following energy function:

$$E(f) = \sum_{\mathbf{p} \in \mathcal{P}} D_{\mathbf{p}}(f_{\mathbf{p}}) + \lambda_{smooth} \sum_{\mathbf{p}, \mathbf{q} \in \mathcal{N}} V_{\mathbf{p}\mathbf{q}}(f_{\mathbf{p}}, f_{\mathbf{q}}), \quad (4.15)$$

where \mathcal{P} is the set of pixels in the image; $D_{\mathbf{p}}$ is the data term that measures how well a planar hypothesis explains a disparity value for a given pixel \mathbf{p} ; $V_{\mathbf{p}\mathbf{q}}(f_{\mathbf{p}}, f_{\mathbf{q}})$ is a smoothness prior computed in a neighborhood \mathcal{N} (in the current chapter a first-order Markov Random Field is considered, thus a neighborhood has four connections); $f_{\mathbf{p}}$ and $f_{\mathbf{q}}$ are the current labels for pixels \mathbf{p} and \mathbf{q} , respectively; and λ_{smooth} is a weighting factor for the regularization term. The $D_{\mathbf{p}}$ function is defined as follows:

$$D_{\mathbf{p}}(f_{\mathbf{p}}) = \begin{cases} \min(C(f_{\mathbf{p}}), C_{max}) & \text{if } f_{\mathbf{p}} \in \{\pi_1, \pi_2, \dots, \pi_n\}, \\ 0.9 \cdot C_{max} & \text{if } f_{\mathbf{p}} \in \{\pi_\infty\}, \end{cases} \quad (4.16)$$

the cost assigned to a pixel \mathbf{p} represents its degree of membership to a given plane π_i . This cost is obtained from $C(\mathbf{p}, d)$ (see Eq. (4.1)), where d corresponds to the hypothetical disparity obtained if \mathbf{p} is assigned to the plane π_i ; if certain hypothesis π_i produces an inconsistent d . For instance, a value outside of the searching space, then \mathbf{p} is penalized with a maximum cost C_{max} . Finally, the smoothness term $V_{\mathbf{p}\mathbf{q}}$ is defined as:

$$V_{\mathbf{p}\mathbf{q}}(f_{\mathbf{p}}, f_{\mathbf{q}}) = \nabla(I_{VS}(\mathbf{p})) \cdot \begin{cases} 0 & \text{if } f_{\mathbf{p}} = f_{\mathbf{q}}, \\ d_{max} & \text{if } f_{\mathbf{p}} \text{ or } f_{\mathbf{q}} \in \pi_\infty, \\ d(f_{\mathbf{p}}, f_{\mathbf{q}}) & \text{otherwise,} \end{cases} \quad (4.17)$$

∇ is the gradient of I_{VS} , and $d(f_{\mathbf{p}}, f_{\mathbf{q}})$ is the Euclidean distance between the points \mathbf{p} and \mathbf{q} depending on which planes they belong to. The minimization of Eq. (4.15) assigns every \mathbf{p} in the reference image one planar hypothesis π_i (see Fig. 4.6(a)). Then, from this membership the corresponding disparity value is obtained by computing the intersection of a ray passing through p with the assigned plane π_i . Figure 4.6(b) shows the dense disparity map corresponding to the illustration used as a case study in previous sections.

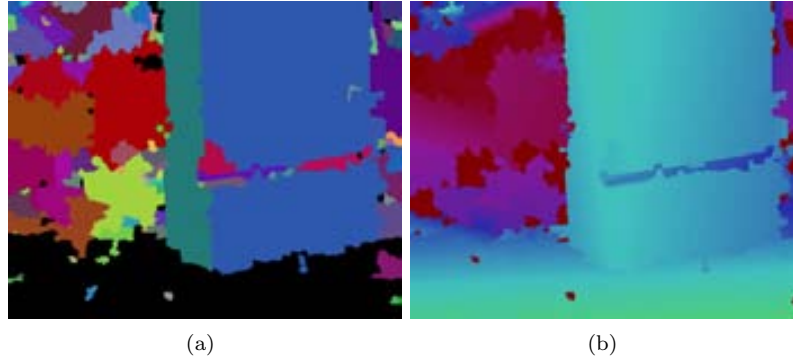


Figure 4.6: Graph cuts outcomes: (a) labelled regions from graph cuts and (b) dense disparity map obtained with the proposed approach.

4.4 Experiments







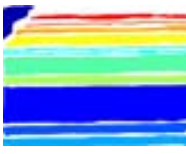

























This section starts presenting some details about evaluation dataset. Particularly, those additional information, which is required for evaluation (ground truth data). Then, eight case studies that provide a qualitative overview of the algorithm performance are presented, these include both intermediate results of the proposed algorithm and dense disparity maps. Finally, two quantitative experiments are conducted over all the evaluation dataset, providing a measuring of global error.

In order to evaluate the proposed approach the dataset presented in Sect. 1.4.1 has been used. This dataset contains 116 scenes, which depicts a large variety of urban scenes, such as: buildings, sidewalk, trees, vehicles, among others (see Table 4.1). Every scene includes their corresponding thermal infrared and color image; a synthesized disparity map; and a hand-annotated map of planar regions. The images are acquired by the proposed multimodal stereo head, and they are processed till to obtain a rectified pair. The planar region maps have been hand-labeled taking into account the geometry of the surfaces, thus a unique label is assigned to each region and it identifies the pixels that belongs to the same plane.

Table 4.1 shows some of the images used for validating the proposed approach. I_{LWIR} and I_{VS} are rectified images; in both cases the size of resulting images is 506×408 pixels. Hand-labeled and disparity maps are given in their original format 640×480 pixels. Since the disparity maps provided by PGB are only accurate in textured regions, we have used a hand-labeled planar regions for obtaining dense and accurate representations, particularly in textureless and noisy regions. To address this problem, we fit a plane to each hand-annotated region, through of image coordinates and corresponding disparity values. The disparity maps resulting from this user supervised labelling process are shown in Table 4.1.

Dense disparity maps were obtained by setting the different parameters as is indicated below. The different values were empirically obtained and the same setting is

Table 4.1: Examples of evaluation data set.

I_{LWIR}	I_{VS}	Maps of planar regions	Synthesized disparity maps
			
			
			
			
			
			
			
			

used in all the scenes. The initial $Dmap_0$ is obtained by defining $d_{min} = 0$ and $d_{max} = 64$. The scale space representation contains three levels and the values used for propagating mutual and gradient information through the different levels: $[\alpha_0, \dots, \alpha_t] = [0.5, 0.3, 0.2]$ and $[\beta_0, \dots, \beta_t] = [0.5, 0.3, 0.2]$; threshold τ_{MGI} is set as 25% of the maximum cost value; finally, mutual and gradient information in Eq. (4.1) are fused defining $\lambda = 0.4$. The two values related with the planar hypothesis generation were set as follow: $\tau_{RANSAC} = 0.2$ and $\tau_{link} = 2.5$. The values given by default in the graph cut implementation provided by [30] were used for the global minimization.

Figures 4.7 and 4.8 show the results obtained for eight different scenes. The initial multispectral stereo images are provided in the first and second columns of Table 4.1. The results are grouped by scenes and they show the output of each step of proposed algorithm. So, every scene has associated four images: (*top-left*) corresponds to support regions R , which split up the I_{VS} image into planar regions; (*top-right*) is an illustration of the planar hypotheses Π ; (*bottom-left*) shows the labelled regions obtained by graph cuts; and (*bottom-right*) is the final disparity map. Notice that Π is the set of labels used during the minimization step, and the disparity map is obtained by using the plane parameters corresponding to each label. On the other hand, it can be appreciate how the minimization stage is able to filter out small regions and propagates information across the neighbors (see (*bottom-left*) illustrations in the different scenes and compare them with their corresponding (*top-right*) images in Fig. 4.7 and 4.8).

Figure 4.1 (scene 5) shows that the proposed approach can obtain dense disparity maps even in non-planar regions. In this illustration a large cylinder is approximated by two planar patches. The number of planar patches depends on the value used for setting the τ_{link} parameter. Even in this challenging case the proposed algorithm is capable of finding a set of planar hypotheses, and converges toward an optimal solution that preserve the appearance of the scene.

The accuracy of proposed algorithm is evaluated by using two metrics. They are frequently employed as quantitative evaluation criteria for stereo matching algorithms. Initially, the absolute mean error (E_{abs}) is computed in a global manner for a given disparity map as follows:

$$E_{abs} = \frac{1}{N} \sum_{j=1}^N |d_C(j) - d_T(j)|, \quad (4.18)$$

where d_C is the disparity map computed by the proposed algorithm, d_T is the ground truth, and N is number of evaluated points. Since, our data set offers reliable ground truth only in those points that lie on a planar region, the error measurement is limited to those image coordinates that have a valid ground truth data and disparity value. Notice that the label π_∞ used during the minimization step corresponds to non disparity, for this reason these image coordinates are excluded from the evaluation. A drawback of using E_{abs} as an evaluation metric lies on the fact that it does not distinguish between few disparity estimations with large errors and lot disparity

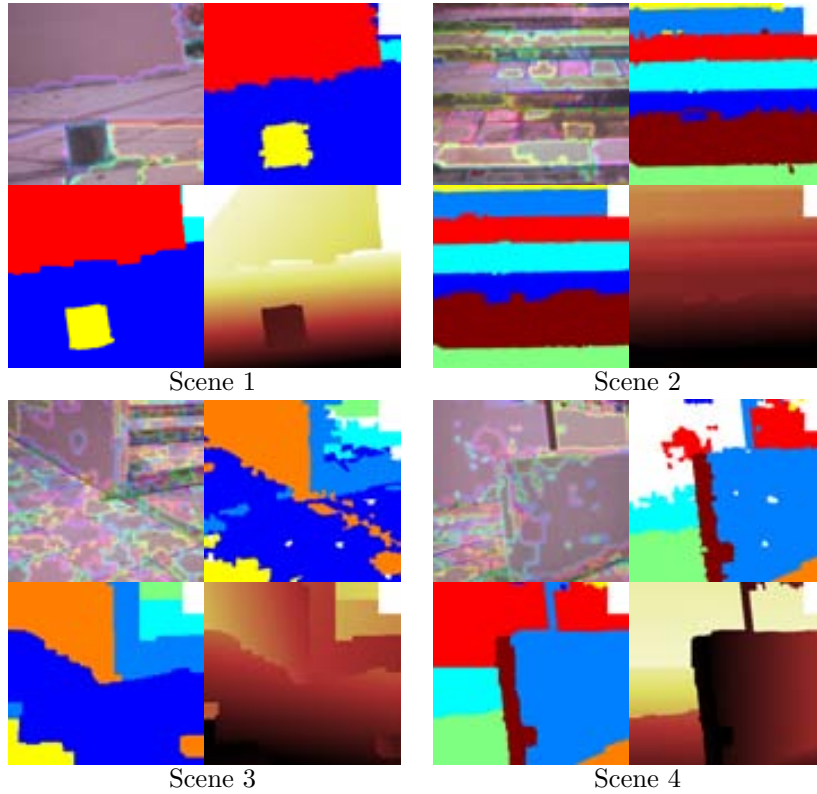


Figure 4.7: Experimental results from different stages of the proposed approach; in each scene the illustrations correspond to: (*top-left*) Support region R ; (*top-right*) Planar hypotheses Π ; (*bottom-left*) Labeled regions from graph cuts; and (*bottom-right*) Final disparity map.

estimations with small errors. Furthermore, it does not take into account that a small disparity value corresponds to a large depth value, and therefore its contribution to the global error should be different, for instance, in comparison to a large disparity (small distance). Hence, in order to take into account this effect, the mean relative error (E_{rel}) is also used. It is computed as follows:

$$E_{rel} = \frac{1}{N} \sum_{j=1}^N \frac{|d_C(j) - d_T(j)|}{d_T(j)}. \quad (4.19)$$

E_{abs} and E_{rel} are computed from the 8 scenes that are used as case studies (see Fig. 4.7 and Fig. 4.8); their corresponding error scores are presented in Table 4.2. The E_{rel} in the scenes 1 and 4 are considerable larger than the rest of scenes in the data set. In both cases these large values result for the wrong matchings due to the lack of texture in the predominant geometries (a vertical non-textured wall).

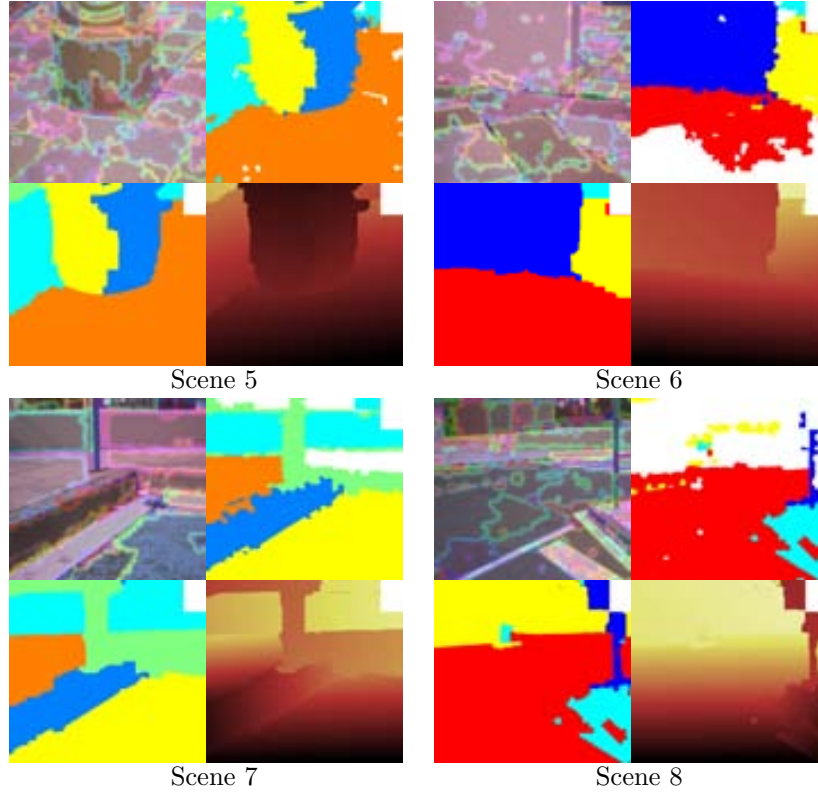


Figure 4.8: Experimental results from different stages of the proposed approach; in each scene the illustrations correspond to: (*top-left*) Support region R ; (*top-right*) Planar hypotheses Π ; (*bottom-left*) Labeled regions from graph cuts; and (*bottom-right*) Final disparity map.

Scene	1	2	3	4	5	6	7	8
E_{abs}	0.635	0.443	0.582	0.629	0.484	0.387	0.465	0.505
E_{rel}	0.167	0.016	0.096	0.144	0.042	0.060	0.062	0.057

Table 4.2: Global E_{abs} and E_{rel} of the case studies presented in Fig. 4.7 and Fig. 4.8.

Figure 4.9 shows the average accuracy of the obtained dense disparity maps, when all the scenes in our data set are considered. For each scene an accuracy histogram is computed by using its corresponding ground truth map. The histogram counts the number of points for a given disparity error, spanning from 0 till 10 pixels. Then, from all these histograms a single plot is computed showing the variability of results (see Fig. 4.9). In this plot the central mark of each box corresponds to its median value and the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme cases.

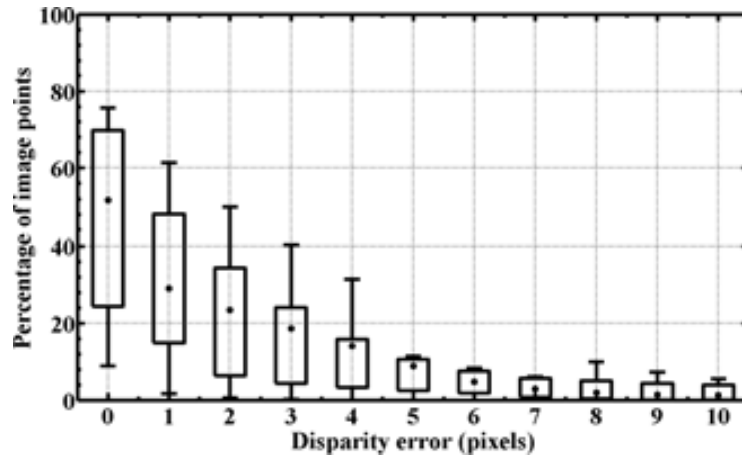


Figure 4.9: Average accuracy of the results obtained with the proposed approach computed from the whole data set.

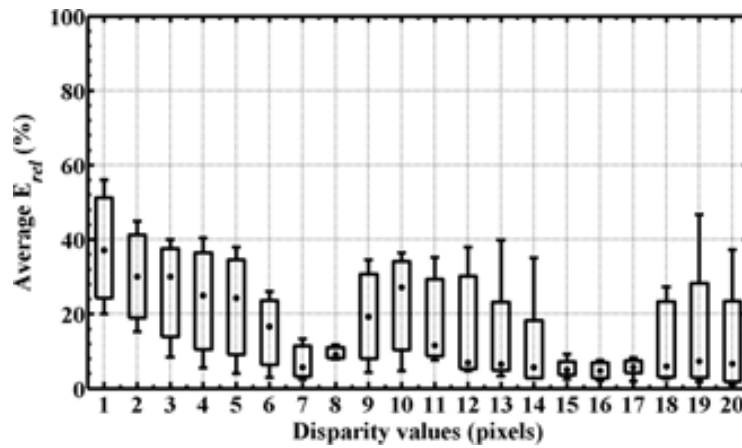


Figure 4.10: Average E_{rel} of the results obtained with the proposed approach computed from the whole data set.

Figure 4.10 shows the average E_{rel} computed from the whole data set. The E_{rel} measurements are restricted to discrete values from 1 to 20 for the sake of visualization. This plot is similar to the previous one and presents the box plot of the mean relative error when all the images into the evaluation data set are considered. Note how the mean E_{rel} decreases to values below 10% when the disparity is higher than 10 pixels. On the other hand, disparity values smaller or equal than 10 pixels correspond to distant points (several meters away from the stereo rig), which are out of the calibration range of the current work.

The results presented above answer the question that was formulated on the beginning (Sec. 4.1), which motivated the current work. They show that under certain restrictions multispectral images can be used to extract dense disparity information. This information can be directly converted into a 3D representation describing the geometry of the scene. This will allow for instance to extract semantic relationships between the objects in the scene.

4.5 Conclusions

The current work presents a novel framework for extracting dense disparity maps from multimodal stereo images, each one of its stages is described as well as the image rectification and camera calibration. The results obtained from this research can benefit those fields where visible and thermal infrared cameras coexist. The main contribution of current work are as follow: (i) it introduces a cost function for obtaining multimodal matching, exploiting mutual and gradient information in a scale space representation; (ii) it proposes a global minimization scheme, which is based on the Manhattan-world assumption, to extract dense disparity maps. Finally, although not a theoretical contribution, a large data set of multimodal stereo images has been generated and is freely available by contacting the authors.

We have shown that under certain restrictions is possible to obtain accurate disparity maps, however the low correlation between thermal infrared and visible images restricts its usefulness in complex environments, being this still an open issue. Future work will be mainly focused on the extraction of a ground truth data, which should includes depth information both of planar and non-planar regions. Additionally, different interest regions such as occlusion and discontinuities would have to be identified, as happen in the (VS/VS) evaluation frameworks for dense stereo algorithm.

Chapter 5

Context-Based Multimodal Stereo

This chapter presents a novel framework for extracting dense disparity maps from a multimodal stereo head using contextual information. It is based on the use of context information, which assume a piecewise planar scene model. This assumption implies that the surfaces of the given scene can be fitted through a compact set of predominant planes. The multimodal stereo head is constructed with a thermal infrared and a color camera. It is intended to explore novel stereo matching approaches that will allow the fusion of information from different sensors. The proposed framework consists of the following stages. Firstly, an initial sparse disparity map is extracted by using a multimodal matching cost volume. Then, a set of plane hypotheses is defined to describe the surfaces on the scene. Finally, the previous stages are combined by reformulating the dense disparity computation as a global minimization problem. Experimental results in outdoor scenarios are provided showing the validity of the proposed framework and its assumptions.

5.1 Introduction

Long wavelength infrared sensors (LWIR), also referred in the literature to as thermal sensors [72, 91], have been generally used for infrared thermography. However, nowadays they are becoming a common sources of information for different computer vision applications. For instance, in video surveillance they facilitate the people detection (hot spots) [52, 58, 37] as well as increase the system availability during night-time. In driving assistance, similarly to the surveillance task, the thermal information helps detecting pedestrians and to analysis occupant posture [32, 93]. Finally, early techniques of thermal performance analysis. For instance, of building isolation; in industrial facilities; and materials test just to mention a few [44, 87], are using VS images as human-machine interface for locating damages. It is clear that the coexistence of LWIR and VS sensor is increasing. In the current chapter, we propose to go a step further by fusing the information from these two cameras (LWIR/VS) with contextual

information in order to obtain dense 3D data.

Over the last few years, many computer vision problems have been investigated from a multimodal point of view, assuming that a multimodal architecture could produce better results than only when a single modality is used. In practice, an inappropriate sensor combination may cause worse results, primarily due to uncertainties of data, since it is difficult to distinguishing the accurate information from biased data [36]. Indeed, a multimodal stereo system as the one proposed here, working at non overlapped spectral bands, supposes a complex task. This system should overcome the well known stereo correspondence problem as well as image fusion issues. A starting point towards such a kind of challenging scenarios was given *tackled* in [22, 42, 48], but in unimodal stereo images affected by smooth radiometric differences. Although they do not truly investigate the multimodal problem, their work indicate that an energy minimization framework based on mutual information could provide dense solutions and accurate results in the multimodal case.

Since energy minimization approaches have demonstrated their usefulness in earlier vision problems, in the chapter, we propose to extend these formulations to multimodal stereo matching. Additionally, we propose to include contextual information as a smoothness term. Hence, our study shows a promissory path towards the integration of multiple sensors for recovering tridimensional information.

There are several work in the field of the extraction of 3D data using contextual information [59, 5, 82, 60]. Particularly for VS/VS stereo systems, it has been shown that is possible to produce an accurate 3D reconstruction, similar to the one obtained with a laser range scanner [28], just by imposing planarity constrains [67]. This assumption may seem to be excessively restrictive. However, such scenes can be effectively modelled by the so-called *Manhattan-world assumption* [15]. Similar to previous work, the proposed approach exploits the contextual information provided by man-made structures. Under this assumption, the surfaces in a given scene are considered as a collection of piecewise-planar patches, where normal vectors are oriented in a reduced number of directions. This condition holds true for images taken in environments with certain regularities in its structure, particularly over the edges.

It should be noticed that although a Manhattan-world assumption is imposed, the proposed approach does not assume that the surfaces into the scene are aligned with a Cartesian coordinate system (X, Y, and Z) as is presented in [28, 67, 6, 61]. Indeed, in the current work the surfaces and their orientations are recovered from an initial sparse disparity map. These surfaces are identified and grouped in accordance with dominant orientations on the scene, which are obtained by a spectral clustering [81]. In this way, false plane detections and noisy surfaces are removed.

The main contribution of current work is the formulation of the multimodal stereo matching as a piecewise planar region partition and labelling. This formulation is then solved by the graph cuts algorithm [9]. The proposed approach works as follows. Firstly, a matching cost volume is computed, which is used for obtaining an initial disparity map. Then, a set of regions is obtained by overlapping two different segmentations of the visible image (perceptual grouping [23] and superpixels [57]). Next, planes are extracted from those candidate regions and are used for partitioning the

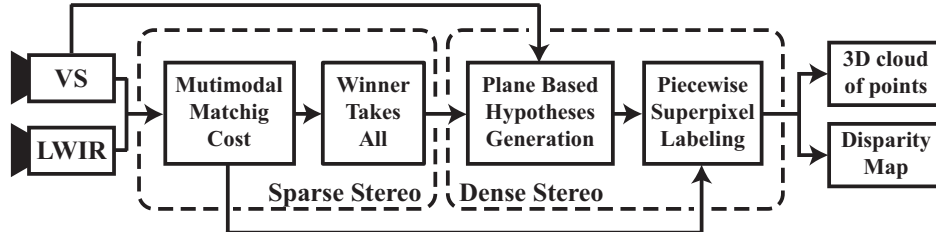


Figure 5.1: Pipeline of the proposed context based multimodal stereo algorithm.

initial disparity map. This map is then treated as a three-dimensional cloud of points; where, the position of each point in that space is defined by its image coordinates (row and column) and disparity value. Once every region is approximated by a plane, the key question is how to obtain the dominant orientations in the scene. Since there is an infinite number of possible orientations, and a plane can be uniquely defined by one point and a normal vector, we propose to tackle this problem as a partitioning problem. Thus, instead of an iterative algorithm that joins regions and computes the planar parameters at each cycle [30], or a progressive approach [96], we propose to group superpixels into clusters, according to their centroid location and normal. The graph consists of a set of nodes that corresponds to all superpixels in the image, and a set of edges whose weights are distances between two superpixels. Finally, that graph is clustered into regions by normalized cuts [81]. We argue that the agglomerations of nodes in the graph mentioned above correspond to dominant orientations in the scene, and in turn they should be the labels used by the energy minimization framework. The proposed clustering strategy has shown high noise immunity, in particular to the noise caused by a deficient plane estimation. Finally, the multimodal energy function is minimized via graph cuts [9]. This function compares at different scales both gradient vectors and mutual information, following a window-based approach. The main steps of the proposed approach are summarized in Fig. 5.1.

The paper is organized as follows. Section 5.2 introduces the state-of-the-art in context based stereo approaches. Then, Section 5.3 details the proposed method. Experimental results are presented in Section 5.4 using several outdoor scenarios showing the validity of the proposed approach. Finally, conclusions are provided in Section 5.5.

5.2 Background

Stereo matching from multimodal images (LWIR/VS) is a field still relatively unexplored, and yet certain questions about image correspondence have not been fully addressed in the literature. By contrast, stereo matching of VS/VS images is an active research area, which evolved from window-based local approaches [46] to global methods that use regularizations for introducing additional information in the form of contextual restrictions or assumptions [86]. The lack of work in the LWIR/VS field

is mainly due to the low correlation between LWIR and VS images, which prevents to compute reliable matching cost values, and at the same time it also prevents to use state-of-the-art methods from VS/VS stereo [42, 48]. A way of avoiding incoherent results and overcoming the correlation problem is by means of the use of local approaches; they commonly reduce the search for correspondences to certain regions of interest (ROI) in the images [43], such as human silhouettes [53], faces [84], or contrasted objects [97] (e.g., hot or cold objects on a smooth background). Although these ROI based approaches have shown attractive results, the main problem remains unsolved. Hence, the search for novel multimodal cost functions is a recurrent subject of research in the multimodal image fusion area. Stereo similarity functions such as Normalized Cross-Correlation (NCC) or image descriptors such as Histograms of Oriented Gradients (HOG) [18] have been evaluated for multimodal stereo matching, but without success [91]. On the other hand, mutual information, Local Self-Similarity (LSS) [80, 90], and the combination of mutual and gradient information in a multiresolution scheme [2] have shown to be the top ranked cost functions [91].

In the current work a mutual and gradient information, in a multiresolution scheme, is used for computing a cost volume that is minimized twice; once with a Winner-Takes-All (WTA) strategy in order to generate a sparse disparity map and then with a graph cut algorithm for a dense representation (disparity and 3D points). Such a kind of method split up the disparity computation in two steps. Initially, a coarse representation of the scene is obtained by a local method that operates at a low level; it is fast and precise. Then, this approximation is refined in subsequent iterations by a method that adds a reasoning layer (high level), whose functionality depends on its complexity. A similar approach has been implemented for the matching of two thermal infrared images in [20, 34, 35]. Although these approaches do not tackle the multimodal stereo matching, they have to deal with similar problems, which are caused by infrared imagery (e.g., low resolution, noise, and blurred edges). They start with a quite sparse representation, which is obtained from feature matching techniques (e.g., corners in [20], and phase congruency of edges in [34] and [35]). Then, this first representation is refined in a second step by removing inconsistencies when a sparse representation is sought [34]; or through the use of support regions such as Delaunay triangulations [20] or watershed segmentation [35] when a dense representation is required. In practice, these methods require a very accurate detection of contours or segmentations since the second step cannot correct mistakes; actually the second step can only reject them.

More general and robust solutions, with respect to the previous methods, have been studied in the VS/VS stereo field; these solutions are usually referred to as the *region-based stereo matching methods*. Early contributions on this topic have tried to associate entities extracted from the images [12]. Entities shall be understood as regions, region descriptors, or nodes of a tree like structure that represents a fine to coarse hierarchical candidate segmentation. A shortcoming inherent of these approaches is the image partitioning, since they need a precise segmentation of both images (homogeneity), which remains as an open issue in computer vision. In contrast, recent approaches have reduced their dependence on the segmentation quality by fitting 3D surfaces, such as planes [6, 50, 88], splines [7], voxels [28], or affine models

[5], to the initial representation. Nowadays, region-based matching are the best ranked algorithms in the VS/VS stereo field (Middlebury¹).

The region based stereo approaches can be classified according to the operation space into two groups: *i*) disparity map based and *ii*) depth map based (e.g., [82], [28], [30] and [7]). Although, the latter has demonstrated a better performance than those using disparity maps, these methods cannot be directly extended to the LWIR/VS stereo matching since the initial depth maps are very sparse. Therefore, in the current work a disparity map based representation is used. The functional structure of the proposed method is not far from the previous reported region based stereo algorithms, as for instance those presented in [6, 30]. The main differences are discussed below.

First of all the proposed approach is based on a split and merge segmentation instead of on a color based algorithm [13, 47] as generally used in region based stereo [96, 6, 50]. Also, we propose to formulate the planar hypotheses generation as a graph clustering problem, instead of a RANSAC-like algorithm [6, 30]. Thus, the phases of plane fitting and planar hypotheses generation are independent, allowing the relaxation of self-similarity assumptions [7]. In the current work, disparity discontinuities between plane patches are measured by an interplane distance as is presented in [88]. This distance is used for grouping plane patches spatially near and with similar plane parameters. Hence segmentation errors can be corrected.

Furthermore, we also do changes in the energy minimization step. Instead of a regular lattice with a neighborhood system of first or second order (4 or 8 connected) [9], a quasi-regular tessellation of sites is used. Each site corresponds to the location of a superpixel centroid and its neighborhood system is limited to those superpixels that share a border. On the one hand, this allows to simplify the labelling problem since the total number of sites is smaller than if a regular lattice were used (number of pixel in the image). On the other hand, a scheme of cost aggregation based on superpixel is introduced, similar approaches have shown a better trade-off between computational efficiency and accuracy in VS/VS stereo evaluations [89].

The energy function used during the minimization step has also been reformulated; the data term is derived from [2] and a novel smoothness term is proposed. This incorporates a piecewise planar prior that penalizes discontinuities between near superpixels, again the interplane distance presented in [88] is used, but it measures the affinity of a pair of planar hypotheses instead of an Euclidean distance as proposed in [30], allowing the minimization over superpixel primitives.

5.3 The Algorithm

A diagram of the proposed approach is shown in Fig. 5.1. The first step corresponds to a sparse stereo, where a cost volume is computed from the LWIR and VS images, and then it is minimized in order to obtain an initial disparity map (hereinafter I_{VS} stands for a VS image and I_{LWIR} for a LWIR image). The second step is a

¹<http://vision.middlebury.edu/stereo/>

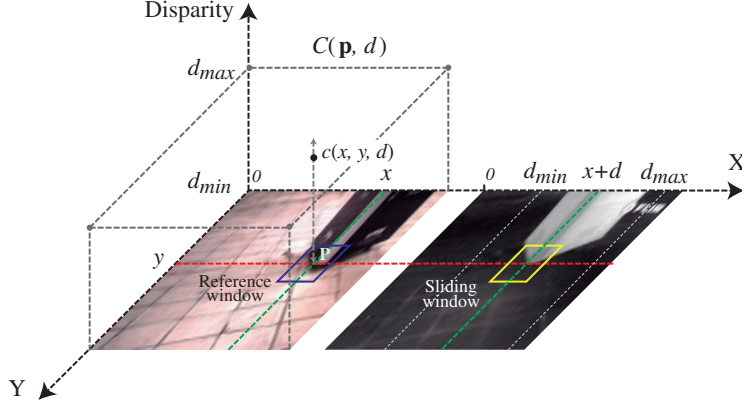


Figure 5.2: Multimodal Matching Cost Volume.

dense stereo algorithm, which starts with the extraction of a set of planes from the initial disparity map. Since labelling problem of plane surfaces is solved by a discrete minimization technique (graph cut), the scene geometry must be summarized into a limited number plane hypotheses. These hypotheses have a double meaning, they are 3D planes with their corresponding geometric parameters, and they are also labels. Once the labelling problem is solved both a dense disparity map and a cloud of points 3D are recovered. In the next three sections, these two steps are explained in more detail.

5.3.1 Multimodal Matching Cost Volume

The multimodal matching cost volume is a three-dimensional array that stores the cost of correspondence between two square windows, obtained from a pair of multimodal images. This volume is obtained following a local window based approach, which consists in computing a cost for each displacement of a sliding window, while a second window is kept fixed on a point in the reference image. The cost volume is referred to as $C(\mathbf{p}, d)$, where $\mathbf{p} = (x, y)$ is the point on reference image (I_{VS}) and d is the disparity or the displacement of the sliding window measured in pixel. The point \mathbf{p} corresponds to the center of the squared window, of size wz , placed on $I_{VS}(x, y)$ whereas d represents the location of the sliding window in I_{LWIR} . Specifically, the latter is a window with the same size than the previous one but centered on $I_{LWIR}(x + d, y)$. Notice that the sliding window location can be parametrized by coordinates of the reference window and d since multimodal images are rectified. Finally, the searching space is defined as an interval $[d_{min}, d_{max}]$ that contains all possible values of d . Figure 5.2 shows how a cost $c(x, y, d)$ is indexed in $C(\mathbf{p}, d)$ together with the windows (i.e., reference and sliding windows) used for its calculation.

The volume computation may be an expensive process, specially if the searching space or the size of the images are large. However, it should be noticed that it is

an efficient representation for two step algorithms since it is computed only once, at the beginning, and then optimized twice (first by a WTA and then through graph cuts). The $C(\mathbf{p}, d)$ is computed using the multimodal cost function introduced in [2], which combines two similarity measures. On the one hand, the mutual information that takes into account pixel values within the two windows, as proposed in [22]. On the other hand, the gradient information that compares gradient vectors within these windows, as proposed in [69]. A multi-scale, or coarse-to-fine scheme, is also included for analyzing the objects in the scene at different resolutions to rise up the matching score [27]. In the current work the elements presented above (mutual information, gradient information, and multiscale scheme) are combined to define the multimodal matching cost volume as follows:

$$C(\mathbf{p}, d) = \lambda C_{MI}(\mathbf{p}, d) + (1 - \lambda) C_{GI}(\mathbf{p}, d), \quad (5.1)$$

where C_{MI} is the mutual information of pixel values, and C_{GI} is the similarity degree of gradient vectors. The λ parameter represents the confidence of MI over GI . In order to increase the discriminative capability of the matching cost function a scale space representation is used. Hence, two stacks of images are generated for each pair of multimodal images, one of them corresponds to a collection of blurred images while the other group contains gradient images (in scale space notation L_0 and L_1 [62]). These representations are obtained by convolving an image (I_{VS} or I_{LWIR}) with a Gaussian kernel of order zero and one, while its standard deviation increases. Figure 5.3 presents a set of images corresponding to a scale space representation². Finally, both MI and GI should be computed at each level t of this hierarchy, and then aggregated into a unique value:

$$C_{MI}(\mathbf{p}, d) = [\alpha_0, \dots, \alpha_t]^T \cdot [MI(\nabla_0^0(I(\mathbf{p}, d))), \dots, MI(\nabla_0^t(I(\mathbf{p}, d)))] \quad (5.2)$$

$$C_{GI}(\mathbf{p}, d) = [\beta_0, \dots, \beta_t]^T \cdot [GI(\nabla_1^0(I(\mathbf{p}, d))), \dots, GI(\nabla_1^t(I(\mathbf{p}, d)))] \quad (5.3)$$

$C_{MI}(\mathbf{p}, d)$ is the resulting cost of propagating mutual information through of the hierarchy, from coarse to fine levels. It is expressed as a linear combination of all values of mutual information for a given position (\mathbf{p}, d) in the stack $\nabla_0^t(I)$ of blurred images, together with a vector of weights that assigns a reliability value to every level. As was mentioned above, the MI operator provides a single value that measures the similarity degree of a pair of windows, considering only the pixel values. The cost from gradient information is treated in an analogous manner.

Mutual information is defined in terms of entropies as:

$$MI(\mathbf{p}, d) = h(\mathbf{p}) + h(d) - h(\mathbf{p}, d), \quad (5.4)$$

where $h(\mathbf{p})$ and $h(d)$ are entropies of reference and sliding windows centered on image coordinate $I_{VS}(x, y)$ and $I_{LWIR}(x + d, y)$ respectively; $h(\mathbf{p}, d)$ is their joint entropy. Mutual information is now formulated as a problem of Probability Distribution Functions (PDF) estimation. Note that it is only necessary to compute $h(\mathbf{p}, d)$, since $h(\mathbf{p})$

² I_{LWIR} images have been contrast enhanced only for the sake of visualization.

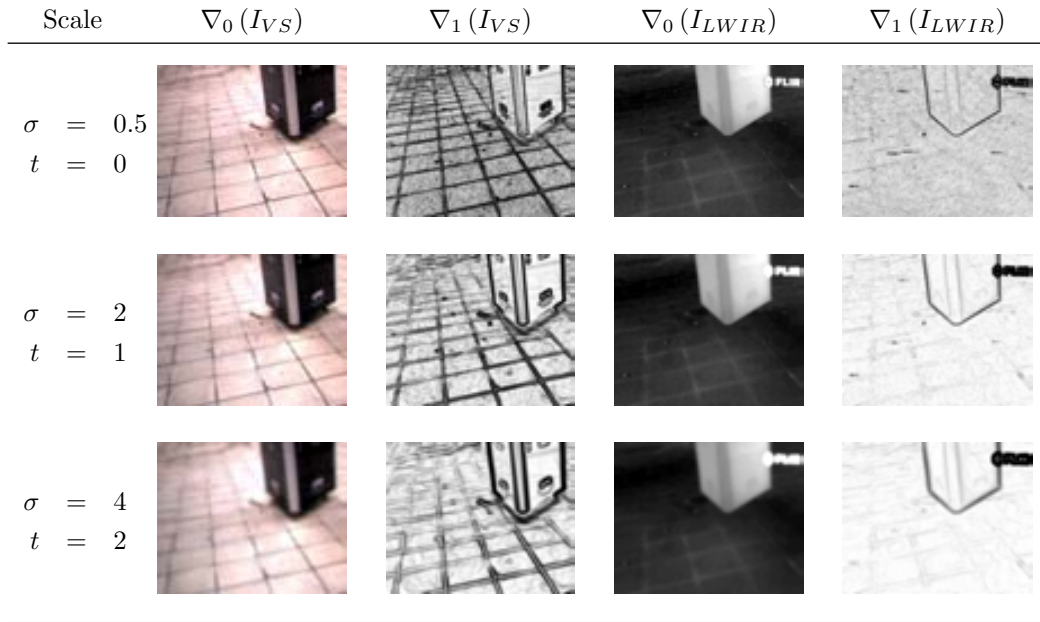


Figure 5.3: Illustration of a set of images defining a scale space representation.

and $h(d)$ are obtained from $h(\mathbf{p}, d)$ [42]. We use a *nonparametric estimator (NP)* [21] for getting the joint PDF $P_{\mathbf{p},d}(i_1, i_2)$. The later is a two dimensional matrix whose cells store the probability that an intensity i_1 corresponds to thermal infrared measuring i_2 . Let us define the joint PDF as:

$$P_{\mathbf{p},d} = NP(\mathbf{p}, d). \quad (5.5)$$

As shown in [48], the entropies in Eq. (5.4) can be estimated by a Parzen window method [94], and expressed as a sum of Gaussian distributions g with standard deviation ψ as follows:

$$h(\mathbf{p}) = - \sum_{i_1} \log(P_{\mathbf{p}}(i_1)) * g_{\psi}(i_1), \quad (5.6)$$

$$h(d) = - \sum_{i_2} \log(P_d(i_2)) * g_{\psi}(i_2), \quad (5.7)$$

$$h(\mathbf{p}, d) = - \sum_{i_1, i_2} \log(P_{\mathbf{p},d}(i_1, i_2)) * g_{\psi}(i_1, i_2), \quad (5.8)$$

where $P_{\mathbf{p}}(i_1) = \sum_{i_2} P_{\mathbf{p},d}(i_1, i_2)$ and $P_d(i_2) = \sum_{i_1} P_{\mathbf{p},d}(i_1, i_2)$ are the sum along each dimension of $P_{\mathbf{p},d}$.

The gradient information is computed from $\nabla_1(I_{VS})$ and $\nabla_1(I_{LWIR})$, by comparing norm and orientation of gradient vectors. That comparison is performed by

couples, taking the gradient vector of a pixel \mathbf{x} and its corresponding \mathbf{x}' ; the final value of GI is computed as the contribution of all these single values as:

$$GI(\mathbf{p}, d) = \sum_{\mathbf{x}, \mathbf{x}'} w(\theta(\mathbf{x}, \mathbf{x}')) \cdot \min(|\mathbf{x}|, |\mathbf{x}'|), \quad (5.9)$$

where θ is the phase difference between two gradient vectors; $w(\theta)$ is a continuous function that penalizes those θ out of phase or counter phase, this is defined as $w(\theta) = (\cos(2\theta) + 1)/2$; finally, $|\mathbf{x}|$ and $|\mathbf{x}'|$ are the norms of these gradient vectors, see [69, 3] for further details.

Once $C(\mathbf{p}, d)$ has been computed, a Winner-Takes-All method is used to select the best disparity for every point in the VS image; then, an initial sparse disparity map ($Dmap_0$) is obtained by filtering unreliable matches using the corresponding matching cost value ($C(\mathbf{p}, d) > \tau$). This minimization step is performed following the procedure presented in [3].

5.3.2 Plane based Hypotheses Generation

The Plane based Hypotheses Generation consists of three steps, which result in a compact set of planar representations that will be used as labels in the final optimization. The first step consists in segmenting I_{VS} into a set of meaningful regions. Since we are working with piecewise planar scenes, ideally, each region will correspond to a plane. Then, in the second step, a plane is fitted to each one of the regions previously obtained, using the sparse disparity map computed in Section 5.3.1. Finally, in the third step, the large set of planes previously computed is compressed resulting in the dominant planes of the scene. The proposed approach is not limited to I_{VS} segmentation; however, segmentation algorithms for VS images are currently the best ranked in tasks of segmentation near to human perception.

Split and Merge Segmentation

In order to overcome the limited information supplied by the initial disparity map, which prevents a correct detection of planar regions, a strategy for partitioning the images into approximately planar regions is adopted. The algorithm works as follows. Initially, I_{VS} is split up into s_i superpixels [57], which are adjusted to the local structure of the image, while preserve edges. The whole set of superpixels in the given image will be denoted hereinafter as S . Then, I_{VS} is again segmented into p_i regions that somehow capture perceptual aspects of the scene [23]. The collection of all p_i is referred to as P . Finally, the superpixels are grouped based on the perceptual regions. The selection of [23] as a merging criterion is due to the fact that the images depict man-made structures, which can be efficiently segmented using an algorithm inspired on perceptual grouping. Furthermore, this algorithm puts special emphasis on edge variabilities, which in the current work is important since it reveals the existence of planar surfaces. The superpixel merging rule is defined as:

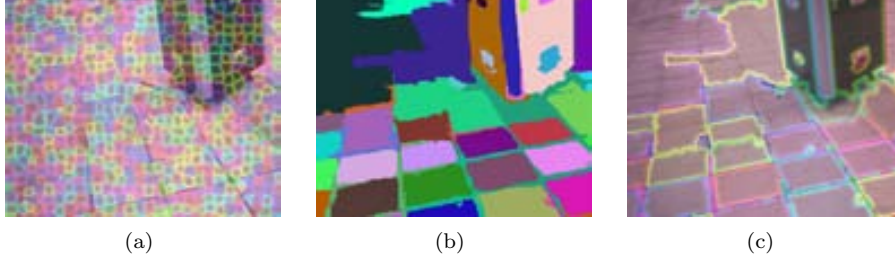


Figure 5.4: Split and merge segmentation example: (a) superpixels (S); (b) perceptual regions (P); and (c) resulting candidate planar regions (R).

$$r_i = \bigcup_{j \in \Omega_i} s_j, \quad \Omega_i = \{j \mid s_j \cap p_i \geq s_j \cap p_k, k \neq i\} \quad (5.10)$$

where Ω_i are the indexes of those superpixels with a maximum overlapping with the given perceptual region p_i . This merging process results in a set of R regions. Figure 5.4 shows an illustration of candidate planar regions R obtained by this split and merge segmentation strategy. This will be later on used to cast the initial disparity map into a lattice-like structure of planar regions. This planar assumption is valid when the number of superpixels is large. In our implementation this value is set to 500, which results in about 150 pixels per region.

RANSAC Plane Fitting

Once the initial disparity map and the candidates planar regions (R) have been obtained (Sections 5.3.1 and 5.3.2), they are combined in order to partition the disparity map into subsets of points; then they are fitted by a plane. Since $Dmap_0$ is considered as a cloud of 3D points (x, y, d) , an estimator of plane parameters based on a Random Sample Consensus (RANSAC) [25] strategy is proposed.

Our estimator alternates between performing a *plane parameters estimation* step, which computes a plane from random samples, and a *plane parameters evaluation* step, which verifies the quality of it. These parameters are then used to describe the corresponding region with a plane. These steps are repeated until a maximum number of iteration is reached or the desired accuracy is obtained. In case that the parameters of a planar region cannot be found, either since the error is greater than the given threshold or the number of iterations is not enough, this region is labelled as non-planar. A more detailed explanation of these steps is given below.

- *Plane parameters estimation.* The parameters (c_1, c_2, c_3) of the plane equation defined as $d = c_1x + c_2y + c_3$, are obtained by randomly selecting three points from the given region—remember that the z coordinate corresponds to the dis-

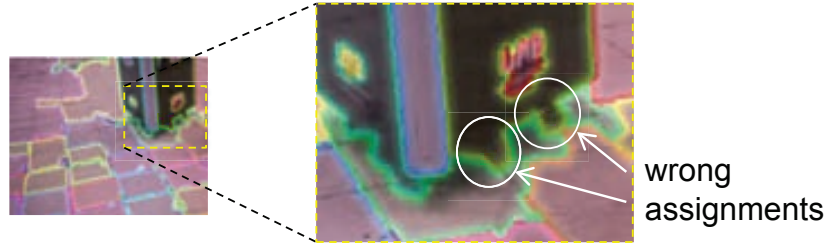


Figure 5.5: Illustrations of superpixels assigned to a wrong region during the split and merge step.

parity value d , while (x, y) to the corresponding column and row position in the image.

- *Plane parameters evaluation.* The quality of the parameters estimated above is evaluated by two criteria: *i*) by counting the number of inliers (i.e., points with a geometric distance to the plane smaller than a given threshold: τ_{ERROR}); and *ii*) by the ratio between number of inliers to total number of points in that region (this ratio needs to be higher than τ_{RATIO}).

The estimator selects the set of plane parameters with the best evaluation (i.e., highest number of inliers and with a ratio greater than τ_{RATIO}). Finally, the selected parameters are refined using only the inliers of the corresponding plane. In this refinement the parameters (c_1, c_2, c_3) are obtained by orthogonal regression using principal components analysis. This RANSAC based plane fitting is repeated with all the regions, resulting in a plane π per region. Hereinafter, each plane is defined by its normal vector \hat{n} and the coordinates of its centroid \bar{x} .

Plane Hypotheses Generation

The previous section could result in a representation that contains as many planes as regions. The current section aims at reducing such a large set into a compact representation, whose members are the most representative planes. The selection of these planes must lead to a compact but also precise set of plane hypotheses that preserve the geometry of scene, since the quality of final results depends on the accuracy of this representation. Although, there are several approaches for constructing such a compact set of plane hypotheses [28, 6, 30], two issues need to be considered here: *i*) how to cluster the planes from the different regions; and *ii*) how the clusters capture the three-dimensional appearance of the scene.

The generation of plane hypotheses is tackled by using a multiclass spectral clustering framework, which is based on graphs and employs *normalized cuts* [81] for weighting the cost of disconnecting two nodes. The proposed clustering approach operates with superpixels. Therefore, each superpixel is represented by a node in the

graph, and inherits the plane parameters of the region r_i that contains this superpixel. The clustering allows to solve wrong assignments generated during the split and merge step (see Section 5.3.2). These wrong assignments are due to the fact that the split and merge step works at pixel level without considering 3D information. They are noticeable in the boundaries of the objects in the scene. Figure 5.5 shows the wrong assignments of two superpixels, which belong to the floor but were merged with the vertical panel during the perceptual grouping.

More formally, we define a graph G as a weighted undirected graph: $G = (V, E, W)$, where V represents the set of nodes (they correspond to the set of superpixels S); E is the set of edges connecting all these nodes; and W is a nonnegative and symmetric distance matrix whose elements corresponds to all the possible distances between the different planes (previously computed by the RANSAC plane fitting approach in Section 5.3.2). The values in W can also be interpreted as a measure of the similarity between two given planes (π_i, π_j) .

As mentioned above, the generation of plane hypotheses is formulated as a graph partitioning problem. Therefore, it is necessary to establish an equivalence relation valid for all pair of nodes, which allows partitioning G into disjoint subsets. Since each subset is considered an equivalence class, all the planes that belong to it are assumed to be similar. In this way, the problem of plane hypotheses generation is simplified to obtain a single element (a plane) to represent each one of the classes. The generation of plane hypotheses works as follows:

1. Given a set of planes (estimated in Section 5.3.2) constructs a weighted graph $G = (V, E, W)$.
2. For a given number of desired clusters computes the normalized cuts on G , as indicated in [81].
3. From the obtained clusters computes the plane parameters (II) using their corresponding clusters' centroid.
4. If the current partition is not enough for encoding the geometry of the scene increases the number of clusters and repeats the iterations from point 2.

The distance matrix W needed to construct the weighted graph G (step (1) in the algorithm) is computed using an inter planar distance d_π , which is formally defined below. This distance allows to establish an equivalence relationship between a pair of connected nodes i and j . $W(i, j)$ is defined as presented in [81]:

$$W(i, j) = e^{-\frac{d_\pi^2(i, j)}{\sigma_{d_\pi}^2}}, \quad (5.11)$$

where d_π is the distance between two planes and σ_{d_π} is its standard deviation. The

distance $d_\pi(\pi_i, \pi_j)$ is obtained, according to [88], as:

$$d_\pi(\pi_i, \pi_j) = l(\pi_i, \pi_j) + l(\pi_j, \pi_i), \quad (5.12)$$

$$l(\pi_i, \pi_j) = \frac{(\bar{x}_j - \bar{x}_i) \cdot \hat{n}_j}{\hat{n}_i \cdot \hat{n}_j}, \quad (5.13)$$

where l is the length of the segment defined by \bar{x}_i and the intersection of \hat{n}_i , passing through \bar{x}_j , with π_j . Figure 5.6 depicts an illustration to make easier the understanding of that distance.

Figure 5.7(a) presents a graph G before the multiclass spectral clustering showing the crowded connections, while Fig. 5.7(b) shows the final clusters obtained after the plane hypotheses generation. It can be appreciated that the latter contains less connections than the former and it is defined by only four predominant planes, which are shown in Fig. 5.7(c).

The generation of plane hypotheses concludes with a set of planes Π whose number is smaller than the initially provided by the RANSAC plane estimator (Section 5.3.2). It is a compact representation of the scene geometry (dominant planes of the scene). Like in the previous case, every plane (Π) is defined by its centroid and normal vector.

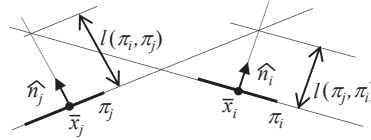


Figure 5.6: 2D Illustration of inter planar distance.

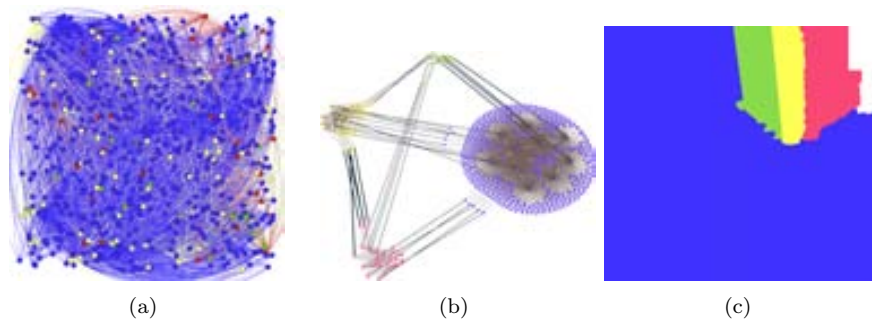


Figure 5.7: Plane hypotheses generation: (a) initial graph G ; (b) resulting partition; and (c) corresponding predominant planes.

5.3.3 Piecewise Superpixel Labeling

Given a compact set of plane hypotheses (Section 5.3.2), the final step consists in performing a piecewise superpixel labeling on reference image I_{VS} . In the current section the set of plane hypotheses computed above is used as labels, and assigned to each superpixel in S . Once every superpixel has been labeled, both a dense disparity and dense depth map are obtained. Dense disparity maps are obtained evaluating the plane parameters associated to a label, while depth maps are obtained from the combination of the disparity with the multimodal stereo head parameters.

The matching of planar regions that belong to different modalities (LWIR / VS) is now formulated as an energy minimization problem in the superpixel domain. Thus, a Markov random field is defined and solved via graph cuts framework [9]. The goal of this section is to obtain a label f that assigns a plane hypothesis to every superpixel. This label minimizes a global energy function \mathbb{E} , which consists of a data term D_s that compares the current label with the observed data, and a pairwise smoothness term V_{st} . This energy function is defined as:

$$\mathbb{E}(f) = \sum_{s \in S} D_s(f_s) + \sum_{s, t \in \mathcal{N}} \lambda_{smooth} V_{st}(f_s, f_t), \quad (5.14)$$

where S is the set of all superpixels; D_s is the data term that measures how well a plane hypothesis explains the disparity value for a given superpixel s ; $V_{st}(f_s, f_t)$ is a smoothness term computed in a surrounding \mathcal{N} of a superpixel; f_s and f_t are the current labels for superpixels s and t respectively; and λ_{smooth} is a constant value used for normalization. Similarly to the work proposed in the VS/VS field [30], in the current work the D_s function is defined as follows:

$$D_s(f_s) = \begin{cases} \min(C_\pi(f_s), C_{max}) & \text{if } f_s \in \{\pi_1, \pi_2, \dots, \pi_n\}, \\ 0.9 \cdot C_{max} & \text{if } f_s \text{ is not a plane,} \end{cases} \quad (5.15)$$

where $C_\pi(f_s)$ is the cost of assigning a plane hypothesis (label f_s) to superpixel s . This cost is defined as follows:

$$C_\pi(f_s) = \sum_{\mathbf{p} \in s} C(\mathbf{p}, d), \quad (5.16)$$

where d is the disparity value obtained by evaluating \mathbf{p} in the current plane hypothesis Π ($d = c_1x + c_2y + c_3$). This cost is equal to the aggregation of costs spanned by the plane f_s in the $C(\mathbf{p}, d)$ volume (Section 5.3.1). Equation (5.15) includes a constant value that is denoted as C_{max} , which is used for: *i*) truncating the C_π cost; *ii*) penalizing inconsistent plane hypothesis for a certain region s (e.g., plane hypothesis that generates disparity values outside of the volume); *iii*) allowing that a given region changes its label by the one of its neighbor.

The smoothness term is defined following [30], but using a superpixel-wise formulation instead of a pixel-wise as initially proposed:

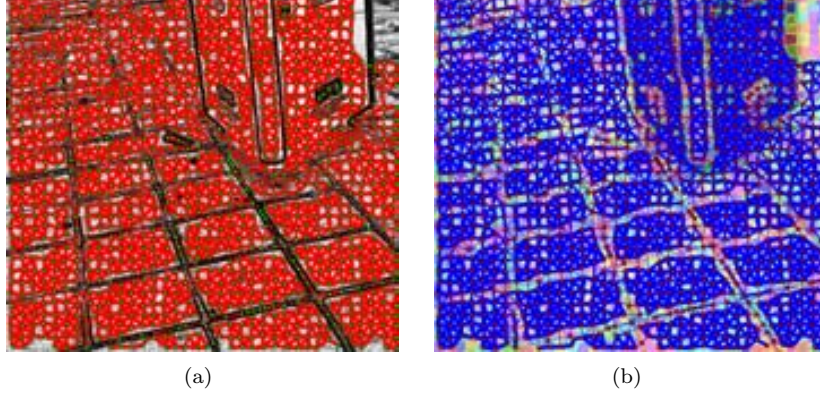


Figure 5.8: Graph connecting neighbor superpixels over: (a) gradient magnitude of VS image; (b) superpixels of VS image.

$$V_{st}(f_s, f_t) = \frac{1}{\underbrace{(\gamma |\nabla_1(I_{VS})| + 1)}_g} \cdot \begin{cases} 0 & \text{if } f_s = f_t, \\ d_{max} & \text{if } f_s \text{ or } f_t \text{ is not a plane,} \\ d_\pi(f_s, f_t) & \text{otherwise,} \end{cases} \quad (5.17)$$

where d_π is defined in Eq. (5.12); d_{max} is a constant value that penalizes discontinuities; g is a weighting function with domain in the gradient magnitude of VS image ($|\nabla_1(I_{VS})|$); the function g is evaluated at the midpoint of the segment that link two superpixel' centroid. Figure 5.8(a) depicts the graph connecting neighbor superpixels—neighbor superpixels are those sharing a common border over $|\nabla_1(I_{VS})|$; the same representation is presented in Fig. 5.8(b) but over the superpixels of I_{VS} . Finally, the energy function defined in Eq. (5.14) is minimized with the graph cut framework presented in [9]. Figure 5.9(a) shows the disparity map of the case study used as an illustration through the manuscript; the corresponding textured 3D map is presented in Fig. 5.9(b).

5.4 Experimental Results

Before presenting the evaluation of results obtained with the proposed approach a brief description of the multimodal stereo system used to acquire the I_{LWIR} and I_{VS} images is presented. Additionally, details about the stereo rig geometry and calibration are also provided.

The multimodal stereo head consists of a pair of cameras (LWIR/VS) separated by a baseline of 12 cm and a non verged geometry. This configuration is obtained by adjusting the pose of the cameras till their z coordinate axis are perpendicular to the baseline. Thermal infrared images are obtained with a *Long-Wavelength InfraRed*

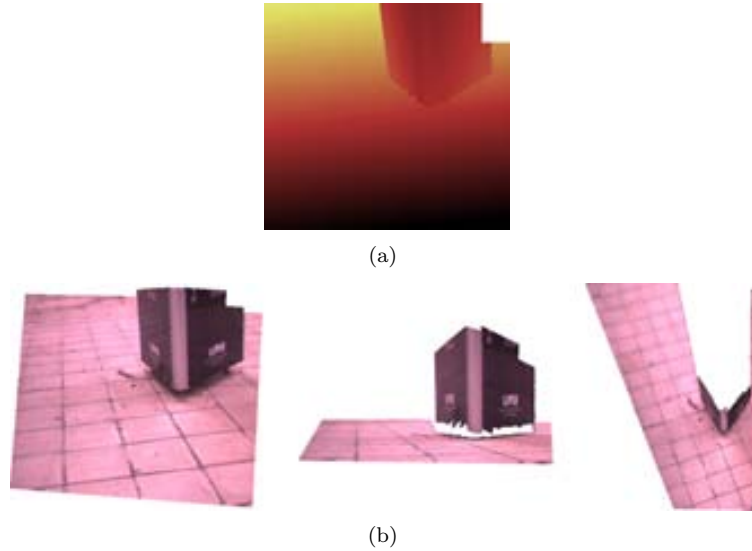


Figure 5.9: Results from the proposed context based multimodal stereo algorithm: (a) disparity map; (b) textured 3D representation.

camera (PathFindIR from Flir³) while color ones with a standard camera based on a ICX084 Sony CCD sensor with 6 mm focal length lens. The former detects radiation in the range of 8 – 14 μm (LWIR band), whereas the color camera responds to wavelengths from about 390 to 750 nm (Visible Spectrum). In order to evaluate the accuracy of the obtained results the proposed multimodal stereo rig is assembled together with a commercial (VS/V_S) stereo head, which is used to generate synthesized ground truth values.

A 2D illustration of the evaluation set-up is shown in Fig. 5.10. This diagram depicts the three cameras, which are linearly arranged, and a point P on the scene. As mentioned above the VS cameras are part of a commercial stereo vision systems⁴ that provides the ground truth data, whereas the multimodal stereo head is composed of the LWIR and VS₁ cameras, as indicated in the figure. The origin of World Reference Frame (WRF), in both stereo rigs (LWIR/V_{S1} and V_{S1}/V_{S2}), is situated on the optical center of V_{S1}, therefore both stereo systems share the same coordinate system.

V_{S1} and V_{S2} are rigidly attached and form a compact device, whose calibration parameters are known. In contrast, the calibration parameters of the multimodal stereo head must be obtained, particularly those related to LWIR camera. These are the intrinsic parameters as well as the rotation (R) and translation (T) with respect to WRF. They are obtained using a standard camera calibration toolbox [8]. Once all the parameters have been estimated, the images are rectified by means of a method proposed in [66]. This method reduces the distortions caused by heterogeneous camera

³www.flir.com

⁴Bumblebee from Point Grey

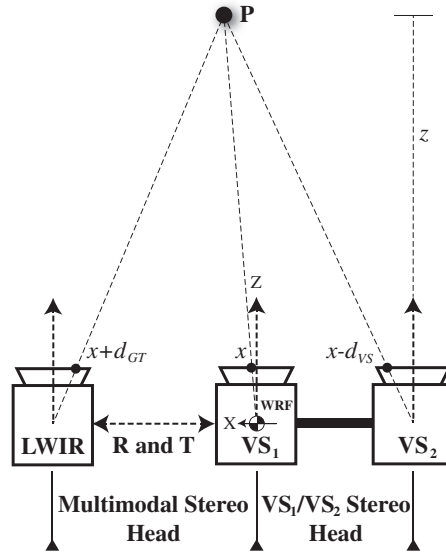


Figure 5.10: 2D representation ((X, Z) plane) of the evaluation set-up.

parameters (LWIR and VS_1).

During the evaluation, these parameters allow to relate outputs of the proposed approach with ground truth data, pixel to pixel, in the following manner. The VS_1/VS_2 stereo head provides both depth and disparity maps. Therefore, z and d_{VS} in Fig. 5.10 are known as well as their projections into I_{VS_1} and I_{VS_2} : x and $x - d_{VS}$ respectively. Since the geometry of the multimodal stereo head has been previously estimated, also the projection of P into I_{LWIR} can be computed from all these values. So, a disparity ground truth data d_{GT} is obtained for every point x at reference image. These values are then used during the evaluation, which is straightforward performed at the level of disparities instead of depths.

Despite the data provided by the VS_1/VS_2 stereo head could be used as a reference, we propose a hand-annotated procedure that helps to improve their precision. The procedure consists of labelling the different planar surfaces in an image. Points with the same label are converted into a planar patch by means of a plane fitting technique (i.e., orthogonal regression using Principal Components Analysis). These planes are used for generating synthetic disparity maps, which are used for the evaluation purpose. Since the evaluation is performed by disparity comparison, a careful labelling leads to better disparity maps, due to noisy data are replaced by approximations and missing disparities values are filled in by interpolations.

The proposed approach has been validated using a large data set that consists of 149 outdoor scenarios. Figure 5.11 shows some of the multimodal stereo images used in both the qualitative and quantitative evaluations of the proposed context based multimodal stereo algorithm. First and second column are rectified images, I_{VS_1} and

I_{LWIR} , respectively; third column depicts their corresponding planar regions (referred to I_{VS1}); finally, the fourth column shows the synthesized disparity maps obtained with the approach presented above.

The images shown in Fig. 5.11 are good examples that fit the Manhattan world assumption; often the images taken in these kind of outdoor environments share properties such as: repetitive patterns (e.g., bricks or floor tile), lack of textured areas due to constant colored regions (e.g., painted walls), or strong lighting changes, just to mention a few. In spite of these drawbacks, Coughlan *et al.* [15] demonstrated that their edge gradient statistics provide information of the orientation of an observer relative to the scene structure, or which objects are not aligned with this structure. We exploit these conclusions by relying on the high-frequency components of the multimodal images. On the one hand, because these high-frequency components (i.e., edges) are the most correlated elements in the images [68]. On the other hand, because disparity values can be efficiently inferred in regions of low-frequency (i.e., non-edges), as VS/VS stereo presented in [28, 67]. The multimodal stereo is more challenging than VS/VS stereo due to both images LWIR/VS must exhibit a Manhattan world. It is not enough that the scene being Manhattan.

Dense disparity and 3D maps have been obtained by setting the parameters as is indicated above. The parameters related to multimodal cost function are obtained following the recommendations presented in [3]. Mutual and gradient information in Eq. (5.1) are fused with $\lambda = 0.45$. The scale space representation has three levels ($t = 2$), the weights in Eq. (5.2) and Eq. (5.3) are set to: $[\alpha_0, \dots, \alpha_t]^T = [0.2, 0.3, 0.5]^T$ and $[\beta_0, \dots, \beta_t]^T = [0.2, 0.3, 0.5]^T$, for C_{MI} and C_{GI} , respectively. These values are obtained maximizing the matching score when the multimodal dataset introduced in [3] is considered. The searching space is limited to a range of $d = \{0, 64\}$, whereas threshold τ is set as 40% of the maximum cost value (see Section 5.3.1). The thresholds used for generating the set of plane hypotheses were set to $\tau_{ERROR} = 0.2$ and $\tau_{RATIO} = 0.6$. Finally, the values used in the global minimization were set as follow: $C_{max} = 5$, $\lambda_{smooth} = 1.25$, $d_{max} = 30$, and $\gamma = 0.5$. The optimal setting of parameters corresponding to the dense stereo step (see Fig 5.1) are computed by means a grid search on a space of parameters, as proposed in [3].

Five case studies are presented in Figures 5.12, 5.13, 5.14, 5.15 and 5.16. They show the experimental results obtained when the scenes shown in Fig. 5.11 are processed. Each illustration corresponds to the outcomes of the different steps of the proposed algorithm. They are: (a) split and merge segmentation; (b) initial disparity map; (c) plane hypotheses generation; (d) graph cut labelling; (e) final disparity map; and (f) different views of the resulting cloud of 3D points. A qualitative inspection of these figures, particularly the illustrations (e) and (f), show that our assumptions are valid to extract dense disparity maps and 3D representations from multimodal images (LWIR/VS).

At this point we consider interesting to give some final remarks about the results presented above. As earlier mentioned, our solution strategy is inspired by region-based stereo matching algorithms, working in the visible spectrum (i.e., [30, 50, 6]). However, in contrast to these methods, a LWIR/VS stereo algorithm must exhibit a

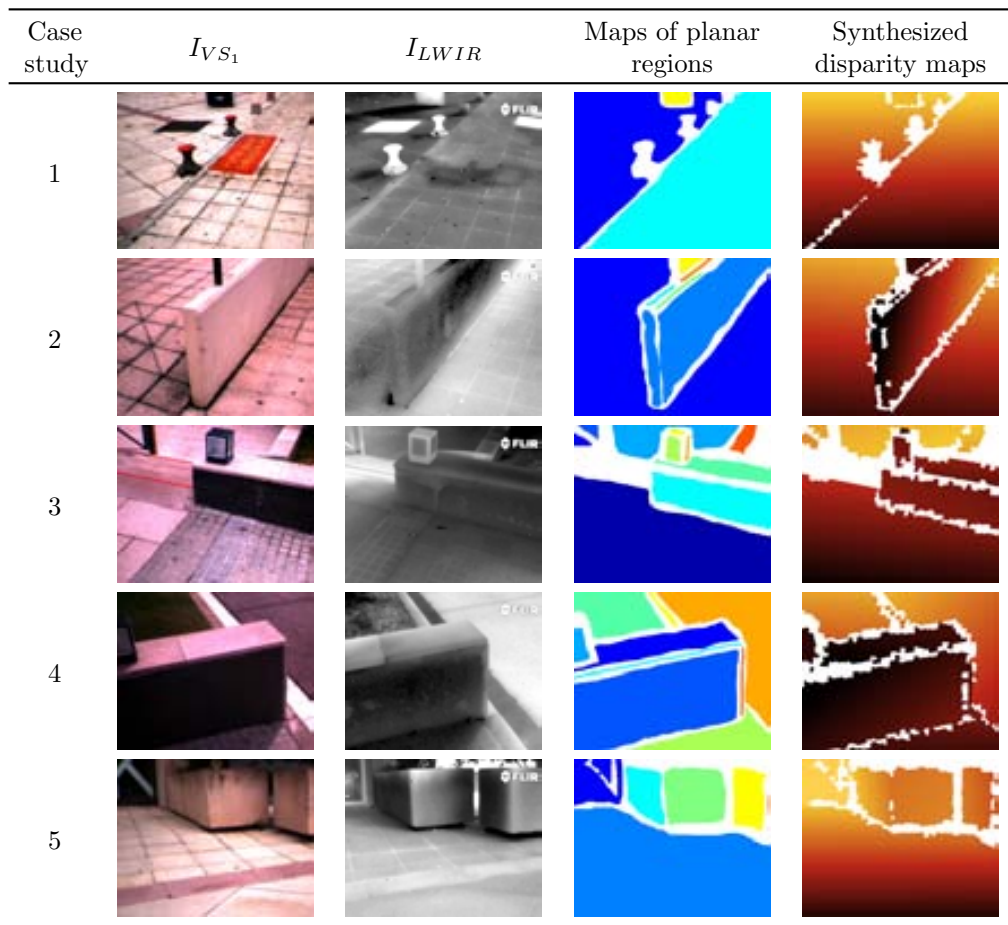


Figure 5.11: Some examples of the evaluation data set consisting of outdoor scenarios containing piecewise planar geometries; both VS and LWIR images are rectified. Images in third and fourth column are aligned to I_{VS_1} .

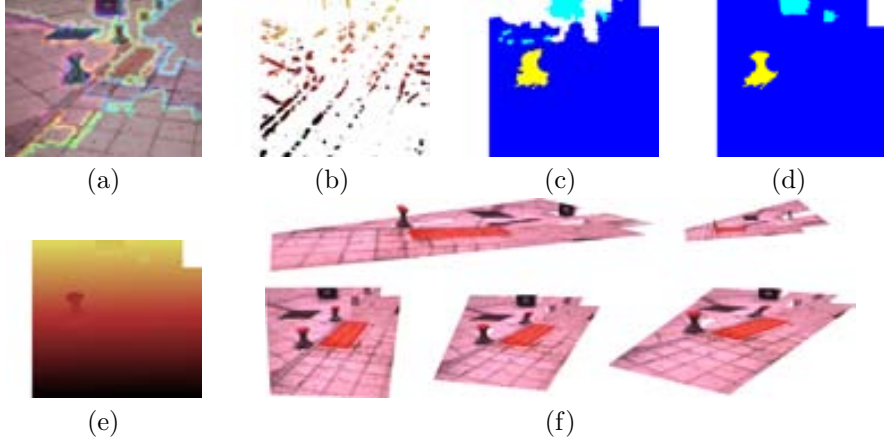


Figure 5.12: Case study 1. (a) Candidate planar regions(R). (b) $Dmap_0$. (c) Plane hypotheses. (d) Regions labelled from graph cut. (e) Resulting disparity map. (f) Different views of the resulting 3D representation.

high noise immunity, which is generated by the uncertainty in the matching costs (data term in Eq. (5.14)). Although this is a common problem to all stereo algorithms, at this particular case, it is a determining factor that may result in poor representations. This drawback is addressed in the current work by assigning more weight to the contribution of the smoothness term, on the contrary to the VS/VIS stereo algorithms mentioned above. This fact can be appreciated in the illustrations (c) and (d) of Figures 5.12 to 5.16, where the right tuning of the weighting parameter allows to filter out small noisy regions and propagate information across the connected superpixels.

The case studies presented in this section validate all the extensions and modification applied over the Markov random field formulation. Furthermore, the disparity maps and 3D representations computed from multimodal images allow to distinguish the most relevant objects in the scene, even if some planes are missed or undetected.

Finally, the performance of the proposed algorithm is quantitatively measured through two error metrics (E_{abs} and E_{rel}). These metrics show the effect of parameters' setting on the obtained results with respect to a synthesized disparity map, which is considered as the *ground truth*. The absolute mean error (E_{abs}) is defined as follow:

$$E_{abs} = \frac{1}{N} \sum_{j=1}^N |d_C(j) - d_{GT}(j)|, \quad (5.18)$$

where d_C is the disparity map computed by the proposed algorithm, d_{GT} is the synthesized ground truth (obtained from VS₁/VS₂ as indicated before, see Fig. (5.10)), and N is the number of evaluated points. Since, the synthesized ground truth is accurate on planar regions, all points out of these regions are excluded from the eval-

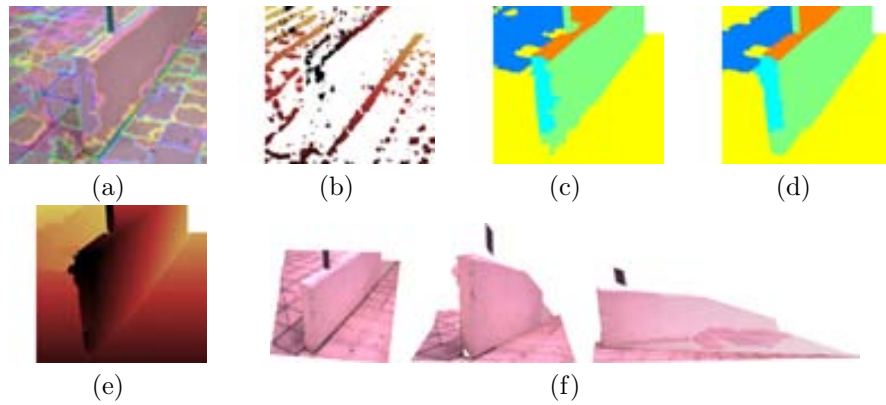


Figure 5.13: Case study 2. (a) Candidate planar regions(R). (b) $Dmap_0$. (c) Plane hypotheses. (d) Regions labelled from graph cut. (e) Resulting disparity map. (f) Different views of the resulting 3D representation.

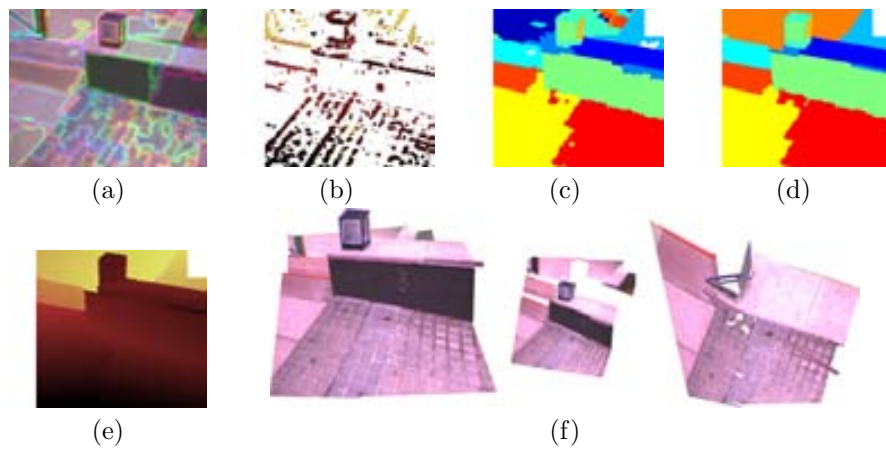


Figure 5.14: Case study 3. (a) Candidate planar regions(R). (b) $Dmap_0$. (c) Plane hypotheses. (d) Regions labelled from graph cut. (e) Resulting disparity map. (f) Different views of the resulting 3D representation.

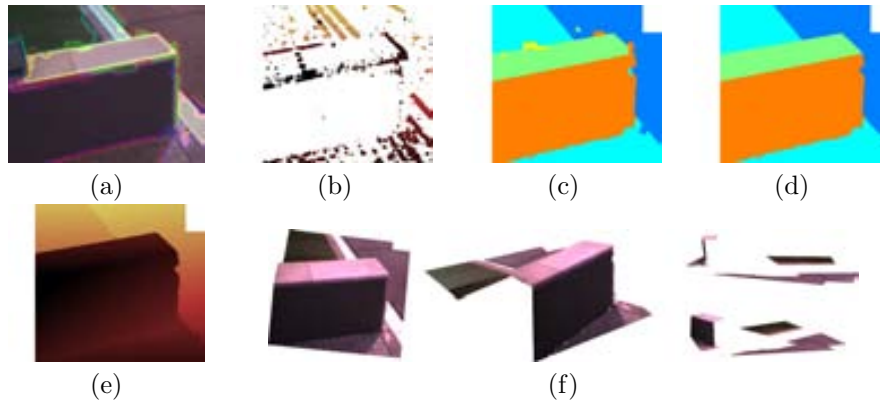


Figure 5.15: Case study 4. (a) Candidate planar regions(R). (b) $Dmap_0$. (c) Plane hypotheses. (d) Regions labelled from graph cut. (e) Resulting disparity map. (f) Different views of the resulting 3D representation.

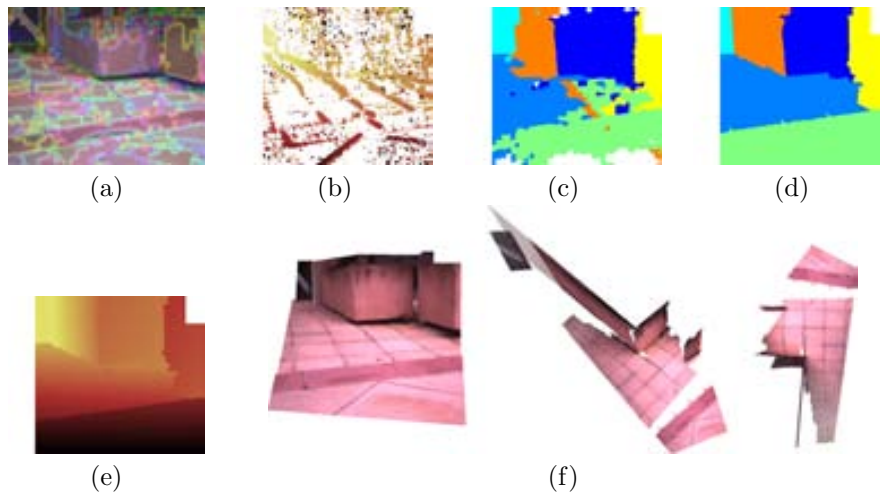


Figure 5.16: Case study 5. (a) Candidate planar regions(R). (b) $Dmap_0$. (c) Plane hypotheses. (d) Regions labelled from graph cut. (e) Resulting disparity map. (f) Different views of the resulting 3D representation.

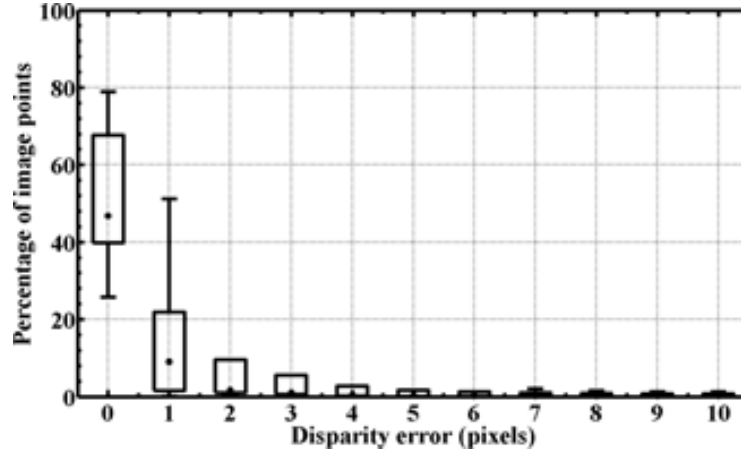


Figure 5.17: Average accuracy of the results obtained with the proposed approach computed from the whole data set (149 pairs of images).

uation process. Moreover, those superpixels are labelled as *not a plane* (Eq. 5.15). The second metric is the relative mean error (E_{rel}), and it is computed as follows:

$$E_{rel} = \frac{1}{N} \sum_{j=1}^N \frac{|d_C(j) - d_{GT}(j)|}{d_{GT}(j)}. \quad (5.19)$$

They are used to evaluate the results from the case studies. They provide quantitative measures of the errors for the different scenes. Table 5.1 presents the E_{abs} and E_{rel} for every case study. On the other hand, Fig. 5.17 shows the average accuracy computed over the whole data set for different disparity error values (in pixels). It is computed as follows. For each image, its corresponding accuracy histogram is obtained, which counts the number of points for a given disparity error. From all these single accuracy histograms the box plot depicted in Fig. 5.17 is obtained.

Case study	1	2	3	4	5
E_{abs}	0.490	0.735	0.572	0.359	0.872
E_{rel}	0.027	0.028	0.020	0.011	0.010

Table 5.1: Global E_{abs} and E_{rel} of the case studies presented in Figures 5.12, 5.13, 5.14, 5.15 and 5.16.

From the quantitative evaluation presented above we can conclude that the proposed approach is able to compute disparity maps with an accuracy of ± 1 pixel in most of the multispectral image pairs of our data set (149 pair of images). Furthermore, it should be highlighted that, on average, about the 40% of matches are correctly detected in every image (disparity error = 0).

5.5 Conclusions

This paper presents a novel multimodal context-based stereo algorithm, which exploits recent advances in VS/VS region-based stereo. It is based on the Manhattan world assumption that allows to split up the given scene in a set of planar patches. These planar patches are then used in a global optimization framework. The main contribution of current work lies on the formulation of the multimodal stereo matching as a piecewise planar region partition and labelling. The experimental results show that the use of the context information helps to overcome the lack of correlation between the multimodal images (LWIR/VS), whereas is a robust way to generate dense scene representations. Furthermore, the proposed approach has shown a noise immunity, in particular in the LWIR images where in general edges are blurred or the image's regions are poorly contrasted. The current work has been tested with a large set of multimodal image pairs showing that context based VS/VS stereo matching algorithms can be extended to tackle the multimodal LWIR/VS stereo problem.

Chapter 6

Conclusions

In the final chapter of this thesis dissertation, we briefly recapitulate the main contributions of our research and discuss possible directions to future work. Finally, publications which are directly related to this thesis are listed in last pages.

6.1 Summary and contributions of this thesis

Chapter 3

This chapter presents a novel multimodal stereo matching algorithm of color and infrared images. The different stages for obtaining sparse depth maps are described. Furthermore, a ROC-based evaluation methodology is proposed for evaluating results from such a kind of multimodal stereo heads. It allows to analyze the behavior over a wide range of different parameter settings. Although the obtained results show a sparse representation, we should have in mind the challenge of finding correspondences in between these two separated spectral bands.

In summary, the main contributions of the current work are: (i) to present a study in an emerging topic as Multimodal Stereo LWIR/VS and achieves a sparse 3D representation from images coming from heterogeneous information sources; (ii) to propose a consistent criteria for making the multimodal correspondence; (iii) to establish a baseline for future comparisons; and (iv) to propose a framework that can be used as a test bed for evaluation purposes in this field.

Next sections will be mainly focused on two aspects: (i) improving the disparity selection process by including Markov Random Fields, which allows to consider prior knowledge of the scene; and (ii) reformulating C_{GMI} as a combination of two individual cost functions, which convert the cost function from a consensus scheme to a scheme where MI and GI contributes to a final matching score according to a set of assignment weights.

Chapter 4

This chapter presents a novel framework for extracting dense disparity maps from multimodal stereo images, each one of its stages is described as well as the image rectification and camera calibration. The results obtained from this research can benefit those fields where visible and thermal infrared cameras coexist. The main contribution of current work are as follow: (i) it introduces a cost function for obtaining multimodal matching, exploiting mutual and gradient information in a scale space representation; (ii) it proposes a global minimization scheme, which is based on the Manhattan-world assumption, to extract dense disparity maps. Finally, although not a theoretical contribution, a large data set of multimodal stereo images has been generated and is freely available by contacting the authors.

We have shown that under certain restrictions is possible to obtain accurate disparity maps, however the low correlation between thermal infrared and visible images restricts its usefulness in complex environments, being this still an open issue. Future work will be mainly focused on the extraction of a ground truth data, which should includes depth information both of planar and non-planar regions. Additionally, different interest regions such as occlusion and discontinuities would have to be identified, as happen in the (VS/VS) evaluation frameworks for dense stereo algorithm.

Chapter 5

This chapter presents a novel multimodal context-based stereo algorithm, which exploits recent advances in VS/VS region-based stereo. It is based on the Manhattan world assumption that allows to split up the given scene in a set of planar patches. These planar patches are then used in a global optimization framework. The main contribution of current work lies on the formulation of the multimodal stereo matching as a piecewise planar region partition and labelling. The experimental results show that the use of the context information helps to overcome the lack of correlation between the multimodal images (LWIR/VS), whereas is a robust way to generate dense scene representations. Furthermore, the proposed approach has shown a noise immunity, in particular in the LWIR images where in general edges are blurred or the images regions are poorly contrasted. The current work has been tested with a large set of multimodal image pairs showing that context based VS/VS stereo matching algorithms can be extended to tackle the multimodal LWIR/VS stereo problem.

Bibliography

- [1] S. Ahmed, T. Hussain, and T. Saadawi. Active and passive infrared sensors for vehicular traffic control. In *IEEE 44th Vehicular Technology Conference*, volume 2, pages 1393–1397 vol.2, Jun 1994.
- [2] F. Barrera, F. Lumbreras, and A. Sappa. Multimodal template matching based on gradient and mutual information using scale-space. In *Proc. IEEE Int'l Conf. Image Processing*, number 10, pages 2749–2752, September 2010.
- [3] F. Barrera, F. Lumbreras, and A. Sappa. Multimodal stereo vision system: 3d data extraction and algorithm evaluation. *IEEE J. Selected Topics in Signal Processing*, 6(5):437–446, sept. 2012.
- [4] E. Bennett, J. Mason, and L. McMillan. Multispectral bilateral video fusion. 16(5):1185–1194, May 2007.
- [5] S. Birchfield and C. Tomasi. Multiway cut for stereo and motion with slanted surfaces. In *Proc. IEEE Int'l Conf. Computer Vision*, volume 1, pages 489–495, 1999.
- [6] M. Bleyer and M. Gelautz. A layered stereo matching algorithm using image segmentation and global visibility constraints. *J. Photogrammetry and Remote Sensing*, 59(3):128–150, May 2005.
- [7] M. Bleyer, C. Rother, and P. Kohli. Surface stereo with soft segmentation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1570–1577, San Francisco, CA, USA, June 2010.
- [8] Jean-Yves Bouguet. Camera calibration toolbox for matlab. http://www.vision.caltech.edu/bouguetj/calib_doc/index.html, July 2010.
- [9] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, November 2001.
- [10] M. Brown and S. Süssstrunk. Multi-spectral sift for scene category recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 177–184, June 2011.

- [11] A. Caballero, J. Castillo, J. Serrano, and S. Bascón. Real-time human segmentation in infrared videos. *Expert Systems with Applications*, 38(3):2577–2584, 2011.
- [12] L. Cohen, L. Vinet, P. Sander, and A. Gagalowicz. Hierarchical region based stereo matching. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 416–421, jun 1989.
- [13] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(5):603–619, May 2002.
- [14] J. Coughlan and A. Yuille. Manhattan world: compass direction from a single image by bayesian inference. In *Proc. IEEE Int'l Conf. Computer Vision*, volume 2, pages 941–947, September 1999.
- [15] J. Coughlan and A. Yuille. The manhattan world assumption: Regularities in scene statistics which enable bayesian inference. In *Proc. Neural Information Processing Systems*, pages 845–851, 2000.
- [16] T. Cover and J. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.
- [17] J. Crowley, O. Riff, and J. Piater. Fast computation of characteristic scale using a half-octave pyramid. In *Proc. Int'l. Conf. Scale-Space theories in Computer Vision*.
- [18] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, volume 1, pages 886–893, june 2005.
- [19] J. Davis and V. Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision Image Understanding*, 106(2-3):162–182, 2007.
- [20] A. Dhua, F. Cutu, R. Hammoud, and S. Kiselewich. Triangulation based technique for efficient stereo computation in infrared images. In *Proc. IEEE Intelligent Vehicles Symposium*, pages 673–678, june 2003.
- [21] N. Dowson, T. Kadir, and R. Bowden. Estimating the joint statistics of images using nonparametric windows with application to registration using mutual information. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30(10):1841–1857, October 2008.
- [22] G. Egnal. Mutual information as a stereo correspondence measure. Technical report, University of Pennsylvania, 2000.
- [23] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *Int'l J. Computer Vision*, 59:167–181, September 2004.

- [24] D. Firmenich, M. Brown, and S. Süsstrunk. Multispectral interest points for rgb-nir image registration. In *Proc. IEEE Int'l Conf. Image Processing*, September 2011.
- [25] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.
- [26] FLIR. Bmw incorporates thermal imaging cameras in its cars lowering the risk of nocturnal driving. Technical report, FLIR Commercial Vision Systems B.V., 2008.
- [27] C. Fookes, A. Maeder, S. Sridharan, and J. Cook. Multi-spectral stereo image matching using mutual information. In *Proc. Conf. 3D Imaging, Modeling, Processing, Visualization, and Transmission*, pages 961–968, 2004.
- [28] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski. Manhattan-world stereo. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1422–1429, Miami, FL, USA, June 2009.
- [29] A. Fusiello, E. Trucco, and A. Verri. A compact algorithm for rectification of stereo pairs. *Machine Vision and Applications*, 12(1):16–22, 2000.
- [30] D. Gallup, J. Frahm, and M. Pollefeys. Piecewise planar and non-planar stereo for urban scene reconstruction. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1418–1425, San Francisco, CA, USA, June 2010.
- [31] T. Gevers and H. Stokman. Classifying color edges in video into shadow-geometry, highlight, or material transitions. *IEEE Trans. Multimedia*, 5(2):237–243, june 2003.
- [32] A. Grassi, V. Frolov, and F. León. Information fusion to detect and classify pedestrians using invariant features. *Information Fusion*, 12(4):284–292, 2011.
- [33] R. Gray. *Entropy and information theory*. Springer-Verlag, New York, USA, 2009.
- [34] K. Hajebi and J. Zelek. Sparse disparity map from uncalibrated infrared stereo images. In *Proc. Canadian Conf. Computer and Robot Vision*, page 17, june 2006.
- [35] K. Hajebi and J.S. Zelek. Dense surface from infrared stereo. In *Proc. IEEE Wksp. Applications of Computer Vision*, page 21, feb. 2007.
- [36] D. Hall and J. Llinas. An introduction to multisensor data fusion. 85(1):6–23, jan 1997.
- [37] J. Han and B. Bhanu. Fusion of color and infrared video for moving human detection. *Pattern Recognition*, 40:1771–1784, June 2007.
- [38] G. Hermosillo and O. Faugeras. Variational methods for multimodal image matching. *Int'l J. Computer Vision*, 50:329–343, 2002.

- [39] H. Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, volume 2, pages 807–814, june 2005.
- [40] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30(2):328–341, February 2008.
- [41] H. Hirschmuller and D. Scharstein. Evaluation of cost functions for stereo matching. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1–8, June 2007.
- [42] H. Hirschmuller and D. Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(9):1582–1599, 2009.
- [43] C. Hua, P. Varshney, and M. Slamani. On registration of regions of interest (roi) in video sequences. In *Proc. IEEE Int'l Conf. Advanced Video and Signal-Based Surveillance*, pages 313–318, july 2003.
- [44] S. Irwan, A. Ahmed, N. Ibrahim, and N. Zakaria. Roof angle for optimum thermal and energy performance of insulated roof. In *icee*, pages 145–150, dec. 2009.
- [45] S. Jung, J. Eledath, S. Johansson, and V. Mathevon. Egomotion estimation in monocular infra-red image sequence for night vision applications. In *Proc. IEEE Wksp. Applications of Computer Vision*, February 2007.
- [46] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: theory and experiment. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 16(9):920–932, sep 1994.
- [47] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24:881–892, 2002.
- [48] J. Kim, V. Kolmogorov, and R. Zabih. Visual correspondence using energy minimization and mutual information. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 1033–1040, October 2003.
- [49] A. Klaus, M. Sormann, and K. Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *Proc. Int'l Conf. Pattern Recognition*, volume 3, pages 15–18, Hong Kong, August 2006.
- [50] A. Klaus, M. Sormann, and K. Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *Proc. Int'l Conf. Pattern Recognition*, volume 3, pages 15–18, August 2006.
- [51] J. Kostlivá, J. Čech, and R. Šára. Feasibility boundary in dense and semi-dense stereo matching. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1–8, June 2007.

- [52] S. Krotosky and M. Trivedi. Mutual information based registration of multimodal stereo videos for person tracking. *Computer Vision Image Understanding*, 106(2-3):270–287, May 2007.
- [53] S. Krotosky and M. Trivedi. On color-, infrared-, and multimodal-stereo approaches to pedestrian detection. *IEEE Trans. Intelligent Transportation Systems*, 8(4), 2007.
- [54] S. Krotosky and M. Trivedi. Person surveillance using visual and infrared imagery. *IEEE Trans. Circuits and Systems for Video Technology*, 18(8):1096–1105, aug. 2008.
- [55] S. Krotosky and M. Trivedi. Registering multimodal imagery with occluding objects using mutual information: Application to stereo tracking of humans. In *Augmented Vision Perception in Infrared*, Advances in Pattern Recognition, pages 321–347. Springer London, 2009.
- [56] A. Kuijper. Mutual information aspects of scale space images. *Pattern Recognition*, 37(12):2361–2373, 2004.
- [57] A. Levinstein, A. Stere, K. Kutulakos, D. Fleet, S. Dickinson, and K Siddiqi. Turbopixels: Fast superpixels using geometric flows. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(12):2290–2297, December 2009.
- [58] A. Leykin and R. Hammoud. Pedestrian tracking by fusion of thermal-visible surveillance videos. *Machine Vision and Applications*, 21:587–595, 2010.
- [59] H. Li and G. Chen. Segment-based stereo matching using graph cuts. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, volume 1, pages 74–81, june 2004.
- [60] Y. Li, Q. Zheng, A. Sharf, D. Cohen-Or, B. Chen, and N. Mitra. 2d-3d fusion for layer decomposition of urban facades. In *Proc. IEEE Int’l Conf. Computer Vision*, pages 882–889, nov. 2011.
- [61] M. Lin and C. Tomasi. Surfaces with occlusions from layered stereo. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(8):1073–1078, aug. 2004.
- [62] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, Norwell, MA, USA, 1994.
- [63] G. Litos, X. Zabulis, and G. Triantafyllidis. Synchronous image acquisition based on network synchronization. In *Proc. IEEE Wksp. Computer Vision and Pattern Recognition*, page 167, June 2006.
- [64] J. Maciel and J. Costeira. A global solution to sparse correspondence problems. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(2):187–199, February 2003.
- [65] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Trans. Medical Imaging*, 16(2):187–198, April 1997.

- [66] J. Mallon and P. Whelan. Projective rectification from the fundamental matrix. *Image and Vision Computing*, 23(7):643–650, 2005.
- [67] B. Micusik and J. Kosecka. Piecewise planar city 3d modeling from street view panoramic sequences. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 2906–2912, june 2009.
- [68] N. Morris, S. Avidan, W. Matusik, and H. Pfister. Statistics of infrared images. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1–7, June 2007.
- [69] J. Pluim, J. Maintz, and M. Viergever. Image registration by maximization of combined mutual information and gradient information. *IEEE Trans. Medical Imaging*, 19(8):809–814, 2000.
- [70] J. Pluim, J. Maintz, and M. Viergever. Mutual information matching in multiresolution contexts. *Image and Vision Computing*, 19(1-2):45–52, 2001.
- [71] S. Prakash, P. Lee, and T. Caelli. 3d mapping of surface temperature using thermal stereo. In *Proc. Int’l Conf. Control, Automation, Robotics and Vision*, pages 1–4, December 2006.
- [72] A. Rogalski. Infrared detectors: an overview. *Infrared Physics & Technology*, 43(35):187–210, 2002.
- [73] N. Salamati, A. Germain, and S. Süsstrunk. Removing shadows from images using color and near-infrared. In *Proc. IEEE Int’l Conf. Image Processing*, pages 1713–1716, September 2011.
- [74] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int’l J. Computer Vision*, 47(1-3):7–42, April 2002.
- [75] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, volume 1, pages 195–202, June 2003.
- [76] Y. Schechner and S. Nayar. Generalized mosaicing: Wide field of view multispectral imaging. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(10):1334–1348, October 2002.
- [77] K. Schreiner. Night vision: Infrared takes to the road. *IEEE Computer Graphics and Applications*, 19(5):6–10, 1999.
- [78] D. Scribner, P. Warren, and J. Schuler. Extending color vision methods to bands beyond the visible. *Machine Vision and Applications*, 11:306–312, 2000.
- [79] S. Shafer. Color. chapter Using color to separate reflection components, pages 43–51. Jones and Bartlett Publishers, Inc., USA, 1992.

- [80] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1–8, june 2007.
- [81] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):888–905, August 2000.
- [82] S. Sinha, D. Steedly, and R. Szeliski. Piecewise planar stereo for image-based rendering. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 1881–1888, Kyoto, Japan, September 2009.
- [83] D. Socolinsky and L. Wolff. Multispectral image visualization through first-order fusion. 11(8):923–931, August 2002.
- [84] D. Socolinsky and L. Wolff. Face recognition in low-light environments using fusion of thermal infrared and intensified imagery. In Riad I. Hammoud, editor, *Augmented Vision Perception in Infrared*, Advances in Pattern Recognition, pages 197–211. Springer London, 2009.
- [85] C. Studholme, D. Hill, and D. Hawkes. An overlap invariant entropy measure of 3d medical image alignment. *Pattern Recognition*, 32(1):71–86, 1999.
- [86] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30(6):1068–1080, June 2008.
- [87] S. Tan, F. Che, Z. Xiaowu, K. Teo, S. Gao, D. Pinjala, and Y. Hoe. Thermal performance analysis of a 3d package. In *Electronics Packaging Technology Conference (EPTC), 2010 12th*, pages 72–75, dec. 2010.
- [88] H. Tao, H. Sawhney, and R. Kumar. A global matching framework for stereo computation. *Proc. IEEE Int'l Conf. Computer Vision*, 1:532–539, 2001.
- [89] F. Tombari, S. Mattoccia., L. Stefano, and E. Addimanda. Classification and evaluation of cost aggregation methods for stereo correspondence. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1–8, june 2008.
- [90] A. Torabi and G. Bilodeau. Local self-similarity as a dense stereo correspondence measure for themal-visible video registration. In *Proc. IEEE Wksp. Computer Vision and Pattern Recognition*, pages 61–67, june 2011.
- [91] A. Torabi, M. Najafianrazavi, and G. Bilodeau. A comparative evaluation of multimodal dense stereo correspondence measures. In *Proc. IEEE Int'l Symp. Robotic and Sensors Environments*, pages 143 –148, sept. 2011.
- [92] P. Torr and A. Zisserman. Mlesac: A new robust estimator with application to estimating image geometry. *Computer Vision Image Understanding*, 78:138–156, 2000.

- [93] M. Trivedi, S. Cheng, E. Childers, and Krotosky S. Occupant posture analysis with stereo and thermal infrared video: Algorithms and experimental evaluation. *IEEE Trans. Intelligent Transportation Systems*, 53(6):1698–1712, November 2004.
- [94] P. Viola and W. Wells. Alignment by maximization of mutual information. *Int'l J. Computer Vision*, 24(2):137–154, September 1997.
- [95] D. Wang and K. Lim. Obtaining depth map from segment-based stereo matching using graph cuts. *Journal of Visual Communication and Image Representation*, 22(4):325–331, May 2011.
- [96] Yichen Wei and Long Quan. Region-based progressive stereo matching. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, volume 1, pages 106–113, June 2004.
- [97] R. Yang and Y. Chen. Design of a 3-d infrared imaging system using structured light. *IEEE Trans. Instrumentation and Measurement*, 60(2):608–617, February 2011.

Publications

The following publications are direct consequence of the research carried out during the elaboration of this thesis, and give an idea of the progression that has been achieved.

Journals

- F. Barrera, F. Lumbreras, A. Sappa. Multimodal Stereo Vision System: 3D Data Extraction and Algorithm Evaluation, In *IEEE Journal of Selected Topics in Signal Processing*, 6(5):437-446, September 2012.
- F. Barrera, F. Lumbreras, A. Sappa. Multispectral Piecewise Planar Stereo using Manhattan-World Assumption, In *Pattern Recognition Letters*, Available online, August 2012.
- C. Aguilera, F. Barrera, A. Sappa, and R. Toledo. Multispectral Image Feature Points, In *sensors*, 12(9), 12661-12672, September 2012.
- F. Barrera, F. Lumbreras, A. Sappa. Context-Based 3D Extraction from Multimodal Stereo Data, In *Information Fusion (Elsevier)*, (current state: second revision).

International Conferences

- F. Barrera, F. Lumbreras, and A. Sappa. Evaluation of Similarity Functions in Multimodal Stereo, In *Proc Int'l Conf. Image Analysis and Recognition (ICIAR)*, pages 320-329, jun. 2012.
- F. Barrera, F. Lumbreras, C. Aguilera, and A. Sappa. Planar-Based Multispectral Stereo, In *Proc. Quantitative InfraRed Thermography (QIRT)*, pages 1-8, jun. 2012.
- C. Aguilera, F. Barrera, A. Sappa, and R. Toledo. A Novel SIFT-Like-Based Approach for FIR-VS Images Registration, In *Proc. Quantitative InfraRed Thermography (QIRT)*, pages 1-8, jun. 2012.
- F. Barrera, F. Lumbreras, and A. Sappa. Multimodal template matching based on gradient and mutual information using scale-space, In *Proc. IEEE Int'l Conf. Image Processing (ICIP)*, pages 2749-2752, sept. 2010.

Technical Reports

- F. Barrera, F. Lumbreras, and A. Sappa. Towards a multimodal stereo rig for ADAS: State of art and algorithms. CVC Technical Report, Computer Vision Center (Universitat Autònoma de Barcelona), July (2008).