Structure from Infrared Stereo Images

by

Kiana Hajebi

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2007

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Kiana Hajebi

Abstract

With the rapid growth in infrared sensor technology and its drastic cost reduction, the potential of application of these imaging technologies in computer vision systems has increased. One potential application for IR imaging is depth from stereo. Discerning depth from stereopsis is difficult because the quality of un-cooled sensors is not sufficient for generating dense depth maps. In this thesis, we investigate the production of sparse disparity maps from un-calibrated infrared stereo images and agree that a dense depth field may not be attained directly from IR stereo images, but perhaps a sparse depth field may be obtained that can be interpolated to produce a dense/semi-dense depth field.

In our proposed technique, the sparse disparity map is produced by a robust features-based stereo matching method capable of dealing with the problems of infrared images, such as low resolution and high noise; initially, a set of stable features are extracted from stereo pairs using the phase congruency model, which contrary to the gradient-based feature detectors, provides features that are invariant to geometric transformations. Then, a set of Log-Gabor wavelet coefficients at different orientations and frequencies is used to analyze and describe the extracted features for matching. The resulted sparse disparity map is then refined by triangular and epipolar geometrical constraints. In densifying the sparse map, a watershed transformation is applied to divide the image into several segments, where the disparity inside each segment is assumed to vary smoothly. The surface of each segment is then reconstructed independently by fitting a spline to its known disparities;

Experiments on a set of indoor and outdoor IR stereo pairs lend credibility to the robustness of our IR stereo matching and surface reconstruction techniques and hold promise for low-resolution stereo images which don’t have a large amount of texture and local details.

Acknowledgements

I would like to sincerely acknowledge my supervisor Professor John S. Zelek for all the guidance, kindness and support he has provided me throughout this study. I cannot imagine completing my thesis without his insightful supervision and continuous support. I would also like to express my appreciation to my thesis readers, Professor Paul Fieguth and Professor Otman Basir, for providing me with their valuable comments.

In addition, I would like to thank the Defense Research and Development Canada (DRDC) for partially supporting this research.

And last, but not least, I would like to thank my family for their never-ending love, support and encouragement along the way.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Goals and Motivation

Infrared (IR) computer vision has recently attracted some interest with the advent of lower cost sensors. The applicability of IR cameras in dark environments is the most motivating factor for using IR imagery in computer vision applications. Some examples include military applications such as target acquisition [20], autonomous vehicle navigation [66], collision avoidance [88], terrain analysis [69], etc.; and surveillance applications like pedestrian detection/tracking [4], [52], face detection/recognition [43], [77], etc.

In many of the aforementioned applications, a required computational vision process is to recover the three-dimensional structure from two-dimensional digital infrared images. This process for visible images is a common feature in many biological vision systems. For example, Human Vision System (HVS) is capable of estimating depth, surface orientation, and spatial relationships, with great accuracy in many situations and circumstances. The main mechanism used by the Human Vision System at distances of less than five meters is stereopsis, which was first described by Charles Wheatstone in 1838 [14]. He discovered that images captured by left and right eyes are different form each other, due to the fact that each eye views the visual world from slightly different horizontal positions. This means that images of objects at different distances from the eyes are projected in the two eyes with different horizontal disparities, giving some cue for depth computation. The same technique has been used in most computational vision applications for three-dimensional structure recovery (e.g., in [4], [21], [27], [45]), usually referred to as stereo computer vision.

The fundamental idea behind stereo vision is to use distances between the positions of a set of primitives in a pair of images with overlapping fields of view. When a very distant object is observed from two cameras positioned in almost the same orientation but separated by a known distance (baseline), that object will appear in relatively similar positions in both images. As the distance between the object and the cameras get smaller, the object's positions in two images will move away from each other. The distance between the projections of a particular object (primitive) in the stereo image pairs is known as disparity; a greater disparity represents a closer object, and a smaller disparity represents an object further away. The most challenging and difficult process in stereo vision is the extraction of the corresponding primitives in two images. Once the correspondences between image primitives have been established, the depth can be easily computed through triangulation.

Computational techniques for extracting depth from stereo has matured significantly and many advances and contributions in computational stereo have been made, allowing stereo to be applied to new types of images such as infrared. However, the potential for discerning depth from stereo using thermal IR imagery has received little attention in the literature. Besides, almost most of the existing techniques for IR stereo matching (like [4], [54], [87]) are based on some assumptions about the hot contents of the images (e.g., humans, faces, etc.) and therefore are not applicable to other situations. This is mainly due to the challenges of this problem, which are introduced and discussed in Chapter 2. The primary objective of this thesis is therefore to develop an efficient method to compute the depth information from stereo images in infrared domain, without having any prior knowledge about the hot or cold objects in the scene.

1.2 Contributions

Methods for computational stereo are generally divided into two groups (according to the type of correspondence algorithm used): *area-based* and *feature-based* [47]. Area-based algorithms match small image windows centered at the primitives that are low level and dense, such as the intensity at each pixel. These techniques result in dense depth maps, but fail when applied to poor textured areas or regions with partial occlusion. In contrast, feature-based algorithms match sparse and more abstract features (such as edges, lines and corners), and provide robust, but sparse disparity maps.

Previous works ([66], [97]) in the area of computational stereo for infrared images, have shown that the quality (i.e., resolution) of infrared sensors is insufficient for calculating dense depth maps. In this thesis, we would like to challenge this result and argue that a dense depth field may not be attained directly, but perhaps a sparse depth field can be obtained that can be interpolated to produce a dense depth field. To this aim, a novel feature-based technique is proposed which involves two phases: (i) *feature matching*, i.e., finding a set of corresponding points in the left and the right images to produce a sparse disparity map; and (ii) *reconstruction*, i.e., producing a dense map from a sparse disparity map.

For the first phase, we present a robust IR stereo matching method, which is composed of three main steps. In the first step, a set of stable and tractable feature points from each image are extracted based on the phase congruency model, which contrary to the gradient-based feature detectors, provides features that are invariant to geometric transformations. We obtain the local frequency information for computing the phase congruency via banks of Log-Gabor wavelets tuned to different spatial frequencies and orientations. The wavelet coefficients are further used in describing and matching the extracted features in the second step. Finally, in the last step, the matching results are further analyzed in order to detect and eliminate the outliers, using the epipolar geometrical constraints.

For the second phase, we develop a surface reconstruction technique to densify the sparse disparity map, obtained from the first phase. Surface reconstruction refers to a process in which a piecewise smooth surface with the same depth discontinuities as the reference image is reconstructed from a set of noisy measurements [13]. Our surface reconstruction method consists of two main steps: (i) extracting the edge map of the reference image (since the edge features provide the exact, non-blurred locations for the discontinuities), and segmenting the image into homogeneous regions, where the disparity can be assumed to vary smoothly inside each region; (ii) approximating the sparse disparity map in each region by a surface interpolation technique.

In addition to presenting a novel technique for depth computation from infrared images, this thesis makes another contribution by performing a comparative study on the efficiency of different interpolation techniques, applied on a set of synthetic and real-world infrared images, in terms of noise and sparseness sensitivity.

1.3 Thesis Outline

The remainder of this thesis is organized as follows. In Chapter 2, we first provide a background on infrared imagery, its characteristics, advantages and challenges of being used in computer vision applications, and then we review the recent works in each of the major applications in this domain. Chapter 3 is started by giving an introductory to stereo processing and a summary of related research in this field. The problem of stereo correspondence is studied by discussing different feature extraction and stereo matching techniques, and then we present our approach to stereo correspondence in detail. The problem of surface reconstruction is discussed in Chapter 4, where we give a brief review on different techniques, evaluate the performance of those methods on synthetic data, and then present the method we adopt to interpolate our sparse disparity map. In Chapter 5, experimental results demonstrating the feasibility of our approach are presented on real world IR stereo images. Finally, in Chapter 6, we draw conclusions about our proposed method and discuss potential future directions for our work.

Chapter 2

Background – IR Computer Vision

In this chapter, we first present an overview of infrared domain characteristics and infrared imaging devices; then the advantages and disadvantages of using infrared sensors over visible spectrum sensors in computer vision-based applications are discussed; and finally an overview of the current and previous works of IR computer vision and IR stereo is presented.

2.1 Infrared spectrum

Infrared radiation is a type of electromagnetic radiation, in which the frequency of waves is higher than that of radio waves but lower than that of visible light. IR radiation is emitted from all objects as a function of their temperature. Hotter objects give off more infrared radiation at higher frequency and shorter wavelength than do cooler objects (see Figure 2.1). Infrared radiation is perceived by humans as heat, when it is emitted from objects at moderate temperatures (above 366K). Unlike visible light, in the infrared domain, every object (even ice cubes) with a temperature above absolute zero emits heat. This leads to the development of Infrared cameras and sensors that can detect IR/thermal energy emitted by the objects in the scene.

Only a certain region of the spectrum is of interest to the IR detectors because much of the radiation in this band is absorbed by water or carbon dioxide in the atmosphere. There are several wavelength bands

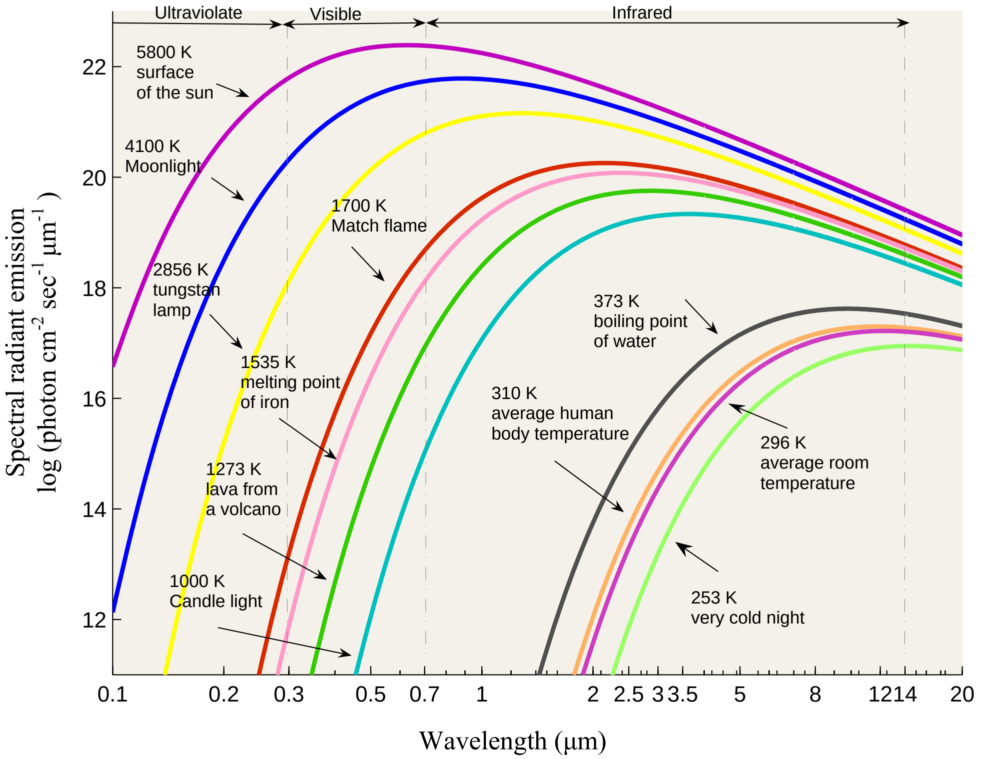


Figure 2.1: Spectral photon emittance at different background temperatures and wavelengths. When the temperature of an object gets higher, the spectral radiant energy at all wavelengths will become higher and the peak wavelength of the emission will become shorter.

with good transmission, that can be detected by IR cameras/sensors[1]:

- **Long wavelength IR (LWIR or far IR) band:** spans roughly 8-14µm, with nearly 100% transmission on the 9-12µm band. The LWIR band offers excellent visibility of most terrestrial objects.
- **Medium wavelength IR (MWIR or MIR) band:** spans roughly 3.3-5.0µm, with nearly 100% transmission, with the advantage of lower background noise.
- **Visible light and short wavelength IR (SWIR or near IR, NIR) band:** spans roughly 0.35-2.5µm, and corresponds to a band of high atmospheric transmission and peak solar illumination, yielding detectors with the best clarity and resolution of the three bands. However, SWIR imagers provide poor or no imagery of objects at 300K (around human body temperature) when no artificial illumination is present.

[1] http://www.xenics.com/

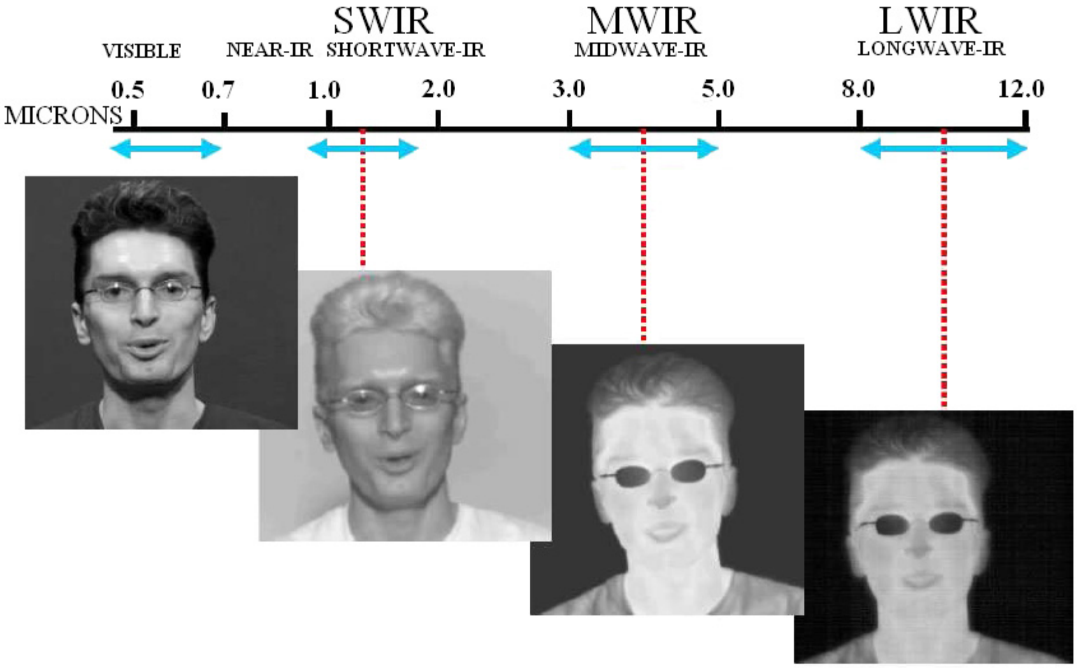


Figure 2.2: The IR images of a human head in different wavelengths. The difference between the far infrared (LWIR) image and the near infrared (SWIR) image shows the independence of reflectance in the former. Furthermore, as can be seen the LWIR and MWIR lights cannot pass through glasses; (Reproduced from [3]).

Figure 2.2 shows some example images of a human face in different wavelengths. Far infrared cameras (which detect LWIR radiation) have been the most widely used infrared sensor for night-vision applications. This is due to the fact that in the long wavelength infrared spectrum, in contrary to the other two bands of spectrum, reflectance of the objects does not contribute to the captured images and only emission from the scene is registered [66]. This has the advantage of eliminating the challenging task of separation of reflectance and emission, which is required for the images captured in the other two spectrum bands. In the remaining of this thesis, the term "infrared" refers to the far infrared spectrum (LWIR).

There are two types of IR/thermal imaging detectors [53]. The first type measures IR indirectly by detecting the changes in heat induced when absorbing IR radiation. Two widely used detectors of this type are the *bolometer* detectors that measure the heat induced electrical resistance change, and the *pyroelectric / ferroelectric* detectors that measure the heat induced electrical capacitance change for certain crystals. The second type of IR imaging detectors is quantum based which operates at low temperatures (e.g., approx 77K) and therefore requires specialized cryogenic cooling. The quantum detectors are very expensive in operation and purchase price, bulky, heavy and power hungry, inconvenient in terms of waiting for pre-cooling before use but have fast response time and ultra-high sensitivity. In contrast, the first type of detectors is less expensive and has slow response time, but operates at high temperatures (i.e., room temperature) and does not need expensive cryogenic cooling.

2.2 Infrared imaging characteristics

One of the most important properties of infrared imagery, that characterizes it from visible imagery, is the ability to operate in environments with poor or no light. This is because the images captured by IR cameras do not vary with changes in lighting conditions but rather change with variation in temperature of the scene objects. Taking advantage of this property of infrared imagery makes it possible to detect people and vehicles in dark environments, which is crucial for many vision applications in the military and surveillance fields. Furthermore, the independence from the scene lighting condition is beneficial for image processing algorithms (particularly in motion related applications), as lighting conditions generally change faster than temperature [32], and besides, no shadow removal is required to be applied on the IR images, since shadows are not captured by IR cameras.

Although the use of thermal imagery alleviates several classic computer vision problems such as shadows, lack of nighttime visibility, sudden illumination changes and etc., IR imagery has its own unique challenges which are briefly explained in the following:

- **IR high noise (low SNR) and low resolution:** in order to use computer vision for IR processing, there are the technical problems of high noise and low spatial resolution that need to be overcome. High noise invalidates the smoothness model that can be a problem for edge detection algorithms based on the differentiation gradient of a smooth 2D surface; Low spatial resolution means losing data (i.e., textures) and losing statistical significance. This makes correspondence based on pixel correlation more difficult, or really impossible, which ironically is necessary for depth from stereo and motion, and is essentially fundamental for 3D vision. This drawback may improve as we get better IR camera technologies.

- **IR reflection:** The surfaces of many materials reflect IR radiation. This reflection is more specular and less diffuse than it is in the visible light. This is mainly due to the longer wavelength of infrared radiation [32]. Examples of IR reflection for a car and a pedestrian standing on a concrete road are shown in Figure 2.3. IR reflection can cause problems in detection systems (particularly region-based ones) as false positive detections may occur in some extreme conditions. Vollmer et al. proposed a solution in [92] that uses a polarizer to partially polarize the specularly reflected light for distinguishing the direct light form the reflected one.

- **IR halo effect or saturation:** Halo around very hot or cold objects which produces a significant contrast with the background is a major characteristic of the thermal images captured by common ferroelectric sensors (see Figure 2.4). The halos[2] have opposite polarity from the objects they surround (i.e., white-black/hot-cold). The presence of halos can cause a serious problem for object detection/extraction techniques, because the halo around an object may be detected as part of the object, and therefore the result does not provide an accurate localization of the object silhouette [18]. However, there are some beneficial consequences of halo effect, as well. Fore example, in pedestrian detection, the presence of halo around pedestrians can help the threshold/contrast-based segmentation methods to detect the pedestrians easier; e.g., in [16], Davis and Sharma proposed a segmentation algorithm which takes advantage of the halo effect. In their method, regions of interest are first segmented using background thresholding. Then inside each region, the precise boundary of the object is extracted by taking advantage of halo effect and examining the high contrast areas.

- **History effects:** This effect is inherent to infrared and cannot be removed by using better IR cameras [53]. The strength of IR radiation depends not only on the current states of the object and the environment, but also on the combined effects of the history of state changes, because temperature variations takes time to propagate and doesn't affect the environment instantly. The brightness constancy assumption, which is the basis for depth

[2] Halo is formed when the infrared radiation is not fully blocked by the mechanical chopper; "a hot source will heat the back of the chopper, and since this secondary radiation is less focused, it will heat the sensor array over a larger area than that of the actual image of the object. It is created when the system electronically subtracts the images with and without chopper obstruction" [32].

from motion (i.e., optical flow), is invalidated by history effect in two ways: (i) the extremely fast change of temperature, which leads to significant brightness variation even between two consecutive video frames; and (ii) a very slow change of temperature like the natural dissipation of heat; this can leave behind a “ghost image” after a hot or cold object moves (that can lead to ghost object detection). An example of this effect is illustrated in Figure 2.5.a, where two adults and a child are shown through an infrared thermal imager. After a minute of sitting on the couch the thermal infrared energy of the people is transferred and stored in the couch, until they get up. The right image illustrates that this energy is now being emitted from the couch and displayed on a thermal imaging device. This information can prove useful in a variety of applications, such as criminal tracking, land/airborne surveillance, and drug facility detection. Figure 2.5.b shows other examples of this effect.

2.3 IR imagery applications

With the rapid advances of infrared sensor technology and its drastic[3] cost reduction, the potential of applications of these imaging technologies in computer vision systems has increased. IR imagery has recently been the focus of many vision research efforts, particularly pedestrian detection/tracking, face detection/tracking, and face recognition. In this section, we briefly review some of the influential works in each of the mentioned applications, in order to illustrate the advantages and the challenges of working with infrared images.

2.3.1 Pedestrian Detection/Tracking

The detection of obstacles and pedestrians in roads is one of the most active research targets in the area of computer vision, due to its important application on road safety. In recent years, night vision systems, exploiting advantages and benefits of IR cameras, have gained more and more interest, for the automatic detection of pedestrians at night ([4], [6], [86]).

[3] Improvements in technology have led to uncooled IR sensors that are smaller in size, use less power, and are lower in price (i.e., approx. $20,000 USD vs. $100,000 USD for the cooled IR sensors that are very expensive in operation and purchase price).

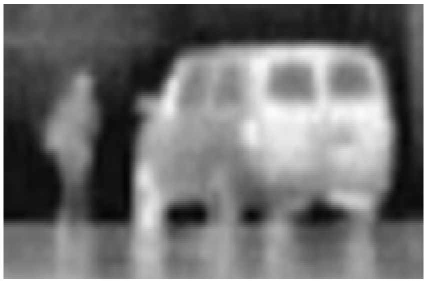


Figure 2.3: Examples of IR Reflection. The reflection of the pedestrian and the vehicle can be seen on the ground; (Reproduced from [32]).

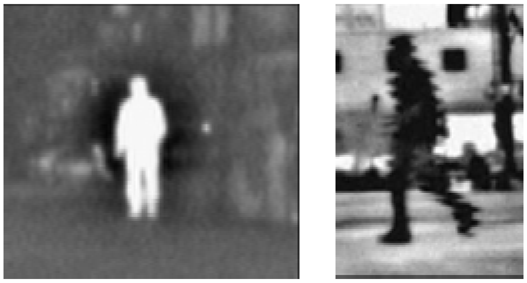
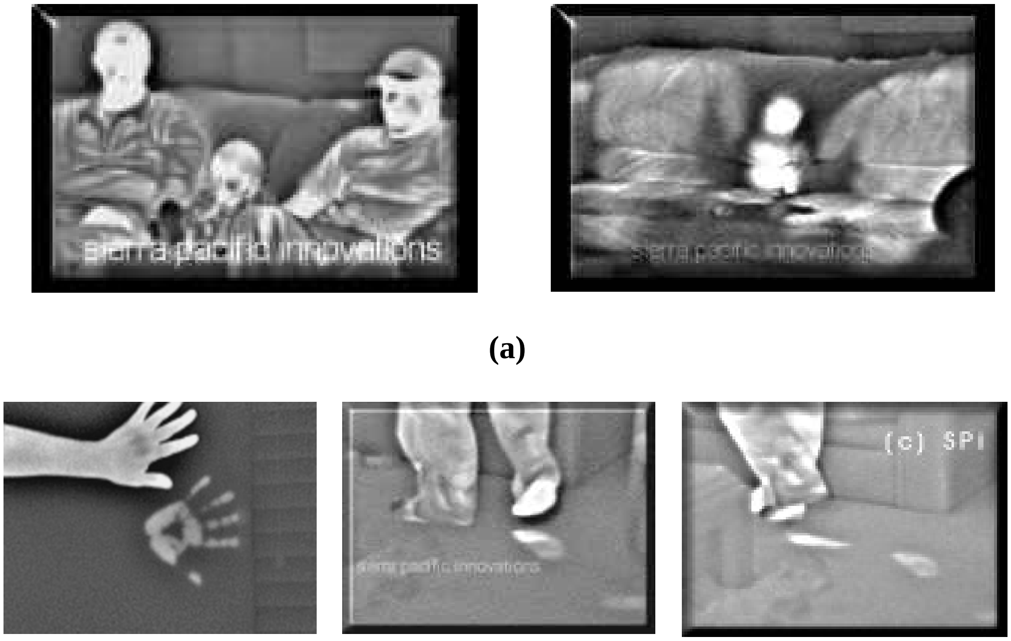


Figure 2.4: Examples of IR halo around the pedestrians; (Reproduced from [32]).


sierra pacific innovations
sierra pacific innovations
sierra pacific innovations
(c) SPI

(a)

(b)

Figure 2.5: Some examples of IR history effect; (a) the thermal energy of people is transferred to the couch after one minute sitting on the coach and is then emitted after getting up; (b) other examples of this effect (Courtesy of Sierra Pacific Corp.).

Although the use of vision sensors and image processing methods provides promising solutions, vision-based pedestrian detection and classification systems must challenge the wide variation in the appearance of pedestrians caused by changes in clothes, carry-ons, posture, and illumination. Infrared imagery provides a good framework for this problem, since the images acquired by infrared cameras only depend on the temperature of objects and the amount of heat they emit, and not the illumination; Furthermore, the temperature/brightness of different pedestrians in different infrared images are roughly similar in spite of different color and textures of their clothing. In addition, infrared cameras are suitable for the detection of objects warmer (or colder) than the environment, due to the sufficient contrast of those objects with the background.

Tsuji et al. [88] have proposed a system to help reducing vehicle-pedestrian accidents occurring at night, using infrared technology. An infrared stereo configuration mounted on a vehicle was used to estimate the distance from objects in the scene. The vehicle's ego-motion, estimated from a gyroscope and the speedometer, was used to capture the relative movement in the environment. A simple threshold-based segmentation was used to locate pedestrian hypotheses in the image, by finding bright regions. The segmented regions are validated based on their sizes, and those are selected with sizes ranging from head-size to full-body size. Distance from the pedestrians is computed by using a correlation-based matching to match the regions around the pedestrians in right and left images. Finally, the relative movement of the vehicle is compared to the pedestrians' movement, and the pedestrians who are on a collision course are reported as the final result. They showed that their algorithm enables judgment of the possibility of a collision 3.5s in advance at cruising speeds of 40-80 km/h.

Bertozzi et al. [4] used an approach based on stereo infrared images to detect areas that are more likely to contain pedestrians. Warm parts of the scene are extracted (assuming pedestrians are brighter than the background) in the right image yielding rectangular bounding boxes around interesting areas. The contents of the resulting bounding boxes are matched with the left image in order to find the corresponding areas, and their relative locations to the camera. The algorithm groups detected objects with similar coordinates, to produce a list of hot areas in which the scene pedestrians are likely to be located. These results are then filtered and only areas with specific size and aspect-ratio are considered and analyzed to find head morphological characteristics. The approach does not use any temporal information, which probably would have improved the performance of the system. The authors analyzed their method on a database created by them, reporting that the system without head detection is able to correctly detect more than 80% of pedestrians in the scene. Enabling the head detection and varying the correlation threshold in the pattern matching, correct detection percentage significantly, reaches the perfect rate (100%) with a very low number of false detection per frame.

Liu and Fujimura proposed a technique [54] as a complementary to previous shape-based approaches [96], [62]. To detect pedestrians, they search for moving objects whose motions are not consistent with the movement of the background. They have developed a two stage stereo correspondence and motion detection procedure that does not compute ego-motion explicitly. Their correspondence method first detects the hotspots (blobs) by thresholding each frame adaptively and then applies a gray-scale template matching for blobs whose counterparts are missing. This method works well in cases where the camera motion has a dominant translational motion with a small amount of rotational motion, which is not always the case. Furthermore, the method makes use of characteristics of night-vision video data, in which humans appear as hotspots; hence it cannot be used for situations where no knowledge about the content of the scene is available.

The techniques which use threshold-based segmentation to locate pedestrians, usually fail to perform well in hot environments (e.g., outdoor environments during summer), where pedestrians are not considerably brighter than the background anymore. Moreover, clothing may mask heat radiation [5]. Therefore, systems based on the assumption that pedestrians are hotter than the environment usually produce partly or complete misdetections. In order to cope with these deficiencies, some pedestrian detection [17], [6] and tracking [52] techniques have been proposed which are based on contemporary use of a visible and an infrared system in order to utilize the benefits of both approaches. Although, only a few research efforts have been carried out on this subject, the result of visible and infrared fusion has been relatively promising.

Davis and Sharma [17] proposed a fusion technique assuming that the two thermal and visible images are co-registered. Using a standard background-subtraction technique, they first identify regions-of-interest (ROIs) in IR images. Color and intensity information within these regions, are then used to extract the corresponding ROIs in the visible image. The input and background gradient information within each region, are combined into a Contour Saliency Map (CSM). The CSM are then thinned and the most salient contours are selected using a thresholding strategy. Contour fragments belonging to the corresponding regions in the thermal and visible images are then fused using the combined input gradient information from both sensors. Any broken contour fragments are completed and closed using a watershed-constrained A* search strategy. Lastly, the contours are flood-filled to produce silhouettes. Their quantitative results on a database created by them, demonstrated that the best performance was obtained by fusing visible and thermal imagery with 76% average sensitivity rate (the fraction of object/person pixels that are correctly detected by the algorithm), and the average sensitivity rate of only-thermal imagery (65 %) was better than the rate of only-visible imagery (43%).

In another work, Bertozzi et al. [6] presented a tetra-vision (4 cameras) system for the detection of pedestrians based on the simultaneous use of one infrared and one visible camera stereo pairs. They processed the two stereo flows independently and then fused the results that come from the different domains, rather than co-registering the images (which seems to be more difficult). The right images of each flow are subdivided into 3×8 pixels regions and their corresponding regions (if available) are detected in the left images. Choosing this size for regions is due to the fact that a pedestrian shape is characterized by strong vertical features (extracted in their method by Sobel filtering). Areas featuring a similar disparity are grouped together and marked by a bounding box. Produced bounding boxes in the visible domain and IR images are then registered and fused together, to produce the final results. The system has proven to be able to detect more than 95% of pedestrians up to 45m and more than 89% up to 75m.

2.3.2 Face Recognition/Detection

Infrared imagery has several advantages over visible imagery, when applied to face detection, detection of disguised faces, and recognition of faces under low illumination and even in total darkness, where visible techniques fail. Face recognition using imaging modalities in spectral bands different from visible (infrared in particular), has become an area of growing interest and attention. As mentioned earlier in this chapter, infrared cameras detect the emitted heat energy and not the reflected light from the objects, and therefore the captured images are independent of illumination, and are less subject to variation by smoke or dust than images captured in the visible domain [43]. The thermal IR images represent the thermal patterns of faces which describe the vein and tissue structure of the faces and are unique for each person (it is known that even identical twins have different thermal patterns [46]). Therefore, the IR images (thermal patterns) can be used to describe and distinguish face images in the task of face recognition.

Face recognition in the infrared-spectrum has received relatively little attention in the literature in comparison with face recognition in visible domain. Several efforts have been made to compare the performance of face recognition techniques in visible and thermal infrared images, and it has been reported that in many cases the performance of thermal face recognition systems is superior to the performance of visible ones (e.g., [76], [77], [44]). In [15], [76] face recognition performance was evaluated using a PCA algorithm on both visible and thermal images. Cutler in [15] reported equivalent performance between mid-wave[4] and visible imagery in an experiment by eigenfaces and a low resolution sensor (a recognition accuracy of 96% was achieved for frontal views, 96% for 45 deg. views, and 100% for profile views). Unlike [15], that kept ambient illumination constant between training and test images, Socolinsky et al. in [76], [77] studies the effect of changing illumination between training and test images, and reported superior performance for thermal infrared imagery (the mean and minimum classification performance of eigenfaces on visible imagery are 78% and 32%; and on LWIR imagery are 96% and 87%, respectively). In [77], they performed a comprehensive comparison of appearance-based algorithms: PCA, LFA, ICA and LDA and they showed that even when illumination is the same for training and test images, thermal infrared recognition performs 4% to 22% better, depending on the algorithm (the weighted mean performances of the aforementioned methods on visible imagery are 73%, 82%, 88% and 93%, respectively; while they are 95%, 93%, 94% and 97%, for IR imagery). In a later work of Socolinsky et al. [78], the verification was also addressed in addition to the identification, and a Monte Carlo approach for performance evaluation was included. Their results indicated that when using visible imagery, the best choice of norms for PCA- and ICA-based recognition yield equivalent identification (94% for visible and 97% for IR) and verification (Equal-error-rate

[4] MWIR images of faces are less illumination invariant than their LWIR counterpart, since the emissivity of skin is lower for MWIR wavelengths.

of 0.09 for visible and 0.06 for IR) performance; performance for LFA-based face recognition are somewhat lower (91% for visible and 96% for IR), whereas LDA-based methods yield the best results (97% for visible and 99% for IR).

In another work**,** Heo et al. [43] proposed a face recognition approach using correlation filters: minimum average correlation energy (MACE) filters and optimum trade-off synthetic discriminant function (OTSDF) filters, and showed that correlation filters have the best performance at low resolutions for both visual (with 84.10% recognition rate) and thermal images (with 96.84% recognition rate), when compared to other face recognition algorithms including PCA, normalized correlation, and LFA. They reported that thermal face recognition shows higher performance than visual face recognition under various lighting conditions and different facial expressions, regardless of which face recognition algorithm is used (with 82.29% for IR and 68.36% for visible, in average).

Other researchers, including [79], [11] have used feature-based approaches (rather than appearance-based methods), to overcome the challenging conditions such as variable poses and facial expressions. Buddharaju et al. in [11] used spectral features. Their method prunes the hypothesis space by modeling the extracted spectral features through Bessel parametric forms. A Bayesian classifier is then used to determine a unique solution from the pruned subset. They showed their method compares favorably to older approaches such as eigenfaces (with over 85% average precision rate, at varying test/training ratios, compared to the precision rate of 78% for other methods).

2.3.3 Other applications

Besides the main applications discussed above, there are some other applications for IR imagery that attracted the attention of different parts of industry. Vehicle navigation is among these applications, where the main challenge is to avoid obstacles at night time, such as those presented in [66], [37]. Owens et al. in [66] developed an unmanned ground vehicle program (Demo III) to enable XUV (Experimental Unmanned Vehicles) autonomous nighttime navigation at speeds of up to 10 m.p.h.. They performed obstacle detection at night and analyzed the suitability of four classes of night vision cameras (3-5µm cooled FLIR, 8-12µm cooled FLIR, 8-12µm uncooled FLIR, and image intensifiers) for night-time stereo vision. Their analysis of signal to noise ratios showed that uncooled infrared sensors lack the necessary sensitivity to produce viable images for nighttime stereo vision during thermal transition periods. (temperature differences on the order of degree produce signal to noise ratio of 40/1 in cooled and 12/1 in uncooled sensors; while the percentage of stereo matches decreases sharply to unacceptable levels for signal to noise ratios below 30/1).

The other approach discussed by Guilloux et al. [37] is to develop a system to help the driver in detecting obstacles and thus allow him a better and earlier reaction in dangerous situations. The system takes advantage of complementary information coming from radar and infrared cameras. Radar is able to tell at what distance points echoes and infrared can give direction to which a relevant event is detected.

Another application of Infrared imagery is occupant posture analysis, such as that presented by Trivedi et al. in [87]. They have developed a real-time vision system for sensing occupant body posture in vehicles and providing safe airbag deployment. They described their experiments on two systems that estimate occupant body position and pose inside a vehicle using long-wavelength infrared (LWIR) imagery and stereo depth data. An edge-based head detection algorithm is simultaneously applied on the visible stereo and IR images, to provide the head pose and size estimates, using the best fit ellipse/head location method. The thermal face detection algorithm remap the IR image to a probability of human skin temperature using a simple Gaussian PDF, with the mean and variance manually, empirically set, and at the same time, the stereo-based face detection obtains the foreground disparity. They showed that although both systems achieve a high success rate (at success rates of 96.4% with visible stereo and 90.1% with IR imagery, for various occupant types), they suffer from certain limitations, like the inability to detect a head location in IR images when the face is obscured with a hat or being turned from the camera and the emissive properties of the subject’s head is changed, making it less elliptical; and the inability in dealing with competing elliptical objects in the scene, especially hands, in stereo images.

2.4 Chapter summary

In this chapter, an overview of infrared domain characteristics and infrared imaging devices has been presented. The advantages and disadvantages of using infrared sensors over visible spectrum sensors in computer vision-based applications has been discussed, and several influential works which used IR imaging for different computer vision applications (e.g., pedestrian detection/tracking, face detection/tracking, and face recognition, occupant posture analysis and vehicle navigation) have been reviewed. Our literature review showed that discerning depth from stereo using thermal IR imagery (which is the topic of this thesis) has received little attention; and besides, almost most of the existing techniques for IR stereo matching are based on some assumptions about the hot contents of the images (e.g., humans, faces, etc.) and compute distance using region matching, which is not sufficiently reliable and accurate, and are not applicable to other situations where no prior knowledge about the content of the scene is available.

Chapter 3

Stereo correspondence for IR images

3.1 Background

The main purpose of computational stereopsis is to reconstruct the 3D geometry of a scene from two (or more) views. The fundamental concept of stereo vision lies in the fact that the projection of a point in three-dimensional world is situated on a unique pair of image locations when observed by two cameras (see Figure 3.1). Consequently, if corresponding points are located in stereo images, then it will be possible to determine the three-dimensional location of the scene point.

The major challenge in computational stereo is *correspondence*. The *correspondence problem* involves searching the images associated with each camera, for the locations that are the projection of the same point in the scene. Because of ambiguous matches (e.g., due to occlusion, specularities, lack of texture) and many-to-many possibilities (e.g., due to similarities of points matched), there exists no general solution to the correspondence problem. Thus, a number of constraints (e.g., epipolar geometry) and assumptions (e.g., image brightness constancy, surface smoothness and proximity) are commonly employed to make the problem more tractable. These constraints and assumptions are explained in Section 3.4.

In the following, first a survey of different correspondence approaches in visual imagery is provided and the drawbacks of using these techniques for IR imagery are discussed. Then we describe our proposed technique for the correspondence problem in the IR domain which

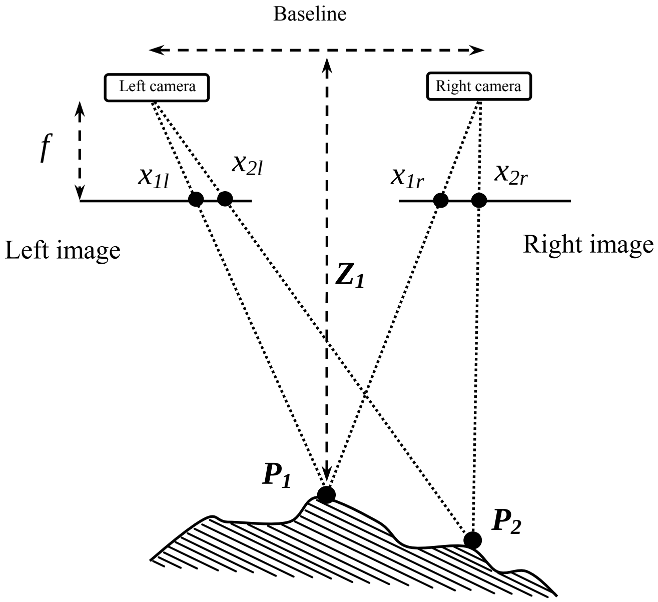


Figure 3.1: Stereo Geometry

consists of three steps:(1) feature extraction, (2) stereo matching and (3) matching refinement.

3.1.1 Stereo correspondence techniques

Solutions to the problem of stereo correspondence can roughly be divided into *global* and *local* methods. In *Global* algorithms, e.g. [45], [9], [28], the disparity assignments are carried out in iterative schemes and on the basis of the minimization of a global cost function. These algorithms result in accurate and dense disparity measurements, though at the expense of higher computational cost. The global function which is minimized in most methods is formulated as [72]:

$$E(d) = Edata(d) + \lambda Esmooth(d)\,. \tag{3.1}$$

The data term, $Edata(d)$, measures how well the disparity function d agrees with the stereo image pair, and $Esmooth(d)$ represents the smoothness assumptions made by the algorithm. The main difference between these techniques is the minimization algorithm used. For example in traditional approaches with regularization and Markov Random Fields, continuation [9] and simulated annealing [28] have been used; and in more recent approaches max-flow [70] and graph-cut [45] have been employed to solve the global optimization problems.

Local stereo matching methods generally fall into two broad categories [19]: area-based (e.g., [59], [38], [65]) and feature-based techniques (e.g., [64], [58]). Area-based algorithms are employed to solve the correspondence problem for every single pixel in the image. They usually use color values and/or intensities within windows of certain sizes. More specifically, for each location in the first image, a rectangular or circular region of pixels around it is matched to the best corresponding region in the second image, using a correlation measure such as cross correlation or sum of squared differences [45]. Although these algorithms have the advantage of producing dense depth fields, they require the images to be highly textures in all parts. Furthermore, choosing the right size for the regions is not easy, as in most cases smaller regions will lead to more mismatches but shorter run-time; while larger regions will produce more accurate results at the expense of higher computational time. Well known algorithms of this type are by Mori et al. [59], Hannah [38] and Okutomi and Kanade [65].

Unlike the area-based techniques, feature-based matching approaches, establish correspondences for feature points only, not all image pixels. These features need to be unique within the left and right images. Examples of good features are: edge points, lines, corners or interest points. Considering that only a small subset of the image pixels are used for matching, these methods result in less detailed depth maps as the depth value is not calculated for every pixel. However, since the possibility of mismatching a feature is less, due to its detailed and distinctive description, feature-based methods usually produce more accurate depth maps. Well known algorithms of this type are Grimson [64], Marr and Poggio [58] and the algorithm of Ohta and Kanade [38] which uses dynamic programming.

Some properties of infrared images introduce challenges to the correspondence problem and limit the performance of most state-of-the-art methods applicable to visible images. As mentioned in Chapter 2, there is the problem of low spatial resolution with infrared images, which makes the correspondence problem based on pixel correlation difficult. Therefore, global and area-based approaches cannot perform well in IR images because the statistical assumption of a sufficiently large amount of data[1] is not satisfied for the low resolution infrared images. On the other hand, since most of the available feature extraction techniques are low level and make little assumption on the underlying imaging modality, they can thus be

[1] Common correspondence techniques depend on the local statistics of regions of texture. When the resolution is not sufficiently high, the number of pixels describing each object is low which makes statistics based method unreliable.

applied to infrared images as well, with little or no modification. Therefore, we chose to use a feature based approach in our problem. Our proposed technique, like the other feature based approaches, consists of three main steps: (i) feature extraction, (ii) feature matching, and (iii) matching refinement. The following sections describe each step of our algorithm in detail.

3.2 Feature extraction

In feature extraction, the goal is to reduce the large number of pixels contained in an image into a set of distinctive primitives. Good features should be stable enough to be repeated in each stereo image, invariant to noise and robust against geometric transformations. Features can be either local (e.g. corners or edges), or global (e.g., polygons or image structures). Global features are difficult and expensive (in terms of computational time) to extract and furthermore they are more application dependent. In contrast, local features are easier to be extracted and in addition, they are more general and are very suitable for situations in which no prior knowledge about the content of the image is available. There exist many robust feature detection methods that have been applied to visible images. As mentioned in the previous section, since most of the available local feature detectors make little assumption on the underlying imaging modality, they can be applied to infrared images as well. However, the relative performance can differ because most infrared images are generally lower resolution and contain more noise than visible-band images. In this section, we first review some of the state-of-the-art feature detectors (e.g. Harris [40], Canny [12], Difference-of-Gaussian (DOG) [56], [57], KLT [85]), and then discuss the feature extraction method we employed for IR images in our system.

3.2.1 Harris corner detection

Among the most popular and widely used feature detectors is the *Harris* operator [40]. The *Harris* operator was explicitly designed for geometric stability. It defines keypoints to be “points that have locally maximal self-matching precision under translational least squares template matching” [90]. In general, these feature points often correspond to corner-like structures. The *Harris* detector searches for pixels (x, y) in the image where the autocorrelation matrix M around (x, y) has two large eigenvalues. The matrix M can be computed from the first derivatives in a window around (x, y), weighted by a Gaussian $G(\alpha = 2)$:

$$M = \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} * G = \begin{bmatrix} A & C \\ C & B \end{bmatrix}, \tag{3.2}$$

where I_x and I_y denote the image gradients in the x and y directions, * is the convolution operator. For any image pixel, where the two eigenvalues (λ_1, λ_2) are large and similar in magnitude, a ‘corner’ is said to occur. To avoid an explicit and computationally expensive eigenvalue decomposition, Harris and Stephens [40] proposed to use a measure based on the determinant and trace of the gradient covariance matrix[2].

$$c(x, y) = \det(M(x, y)) - k \times trace^2(M(x, y)). \tag{3.3}$$

In the above equation, $\det(M) = \lambda_1 \lambda_2 = AB - C^2$ and $trace(M) = \lambda_1 + \lambda_2 = A + B$. The second term is used to eliminate contour points with a strong eigenvalue; the parameter k is usually set to 0.04-0.06. The final result is obtained by applying a non-maxima suppression using a 3×3 mask, and a threshold to select interest points. Figure 3.3.a show the result of this operator on an IR stereo image pair.

3.2.2 Difference of Gaussian (DoG) features

Lowe [56], [57] has proposed a method to identify locations in image scale space that are invariant with respect to image scaling and rotation, and are minimally affected by noise and small distortions. Lowe presented a technique for building up a scale space for an image (*I*), by using a difference-of-Gaussian (DoG) function $D(x, \sigma)$, which can be efficiently obtained from the difference of two adjacent scales that are separated by a factor of *k*:

$$D(x, \sigma) = (G(x, k\sigma) - G(x, \sigma)) * I(x). \tag{3.4}$$

[2] This is a slightly different version from the one that has been first defined by Noble [63]: $c(x, y) = trace(M(x, y)) / \det(M(x, y))$

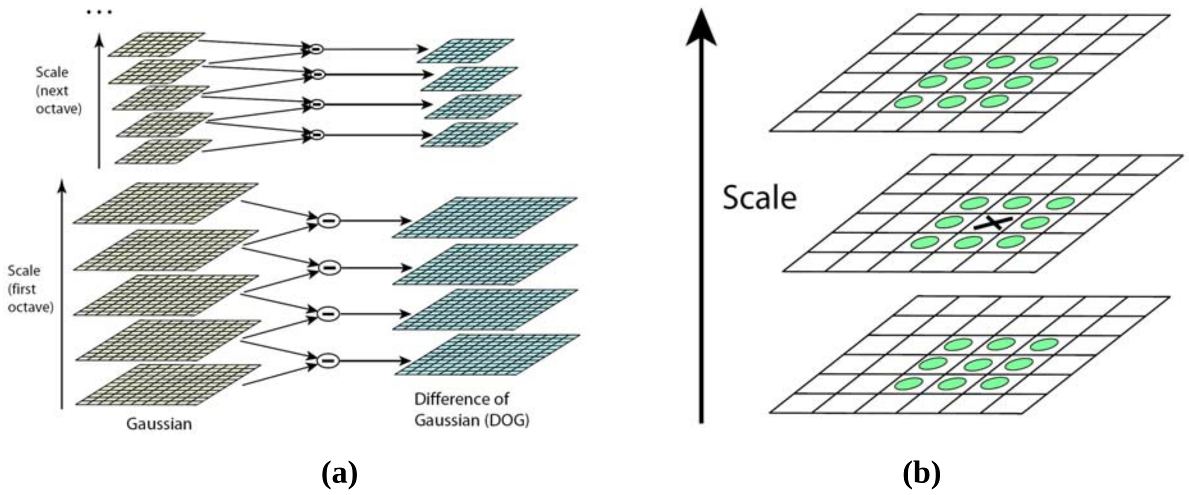


Figure 3.2: (a) The left image represents the result of repeatedly convolving the initial image with a set of Gaussians, for each octave of the scale space. The right image illustrates the difference-of-Gaussian images, produced by subtracting adjacent Gaussians. (b) In order to detect maxima and minima of the difference-of- Gaussian images, each considered pixel (marked by X) is compared to all its neighbours within 3x3 windows at the current and adjacent scales (neighbours are marked with circles); (Reproduced from Lowe [56]).

Lowe [57] shows that when this factor is constant, the computation already includes the required scale normalization. He chooses this factor by dividing each scale octave into an equal number *K* of intervals, such that $k = 2^{1/K}$ and $\sigma_n = K^n \sigma_0$.

For an efficient computation, the resulting scale space can be implemented with a Gaussian pyramid, which re-samples the image by a factor of 2 after each scale octave. The DoG interest points are defined as locations that are simultaneously extrema in the image plane and along the scale coordinate of the $D(x,\sigma)$ function. Such points are found by comparing the $D(x,\sigma)$ value of each point with its 8-neighbourhood on the same scale level, and with the 9 closest neighbours on each of the two adjacent levels (see Figure 3.2). Since the scale coordinate is only sampled on discrete levels, it is important to interpolate the responses at neighbouring scales in order to increase the accuracy of detected features locations. In practice, this is done by fitting a second order polynomial to each candidate point and its two closest neighbours (see [57] for more details). Finally, those points are kept that pass a threshold t and whose estimated scales fall into a certain scale range $[s_{\min}, s_{\max}]$. Figure 3.3.b shows the results of applying this detector on two sample images.

3.2.3 Canny edge detection

Among the most important and powerful edge detectors, which has been widely used in computer vision applications including stereo correspondence, is the edge detector developed by Canny [12], which detects the edge points and simultaneously suppresses the noise using the following algorithm:

1. Smoothing the image I with a Gaussian filter G_σ in order to reduce noise:

   $$J(i, j) = G_\sigma(i, j) * I(i, j) \ , \tag{3.5}$$

   where

   $$G_\sigma(i, j) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{i^2 + j^2}{2\sigma^2} \right] . \tag{3.6}$$

2. Computing the gradient components of $J(i, j)$ using any gradient operator (e.g., Roberts, Sobel, etc), $\nabla J = [J_x, J_y]^T$, to estimate the edge strength:

   $$e_s(i, j) = \sqrt{J_x^2(i, j) + J_y^2(i, j)} \ , \tag{3.7}$$

   and the orientation of the edge normal:

   $$e_o(i, j) = \arctan \frac{J_y}{J_x} . \tag{3.8}$$

   The output is the strength image E_s, created by the values $e_s(i, j)$, and an orientation image E_o, created by the values $e_o(i, j)$.

3. Threshold E_s:

   $$E_s(i, j) = \begin{cases} E_s(i, j) & if \quad E_s(i, j) > T \\ 0 & otherwise \end{cases} , \tag{3.9}$$

   where T is so chosen that all edge elements are kept while most of the noise is suppressed.

4. Suppress non-maxima pixels in the edges in E_s obtained above to thin the edge ridges (as the edges might have been broadened in step 1). To do so, check to see whether each non-zero $E_s(i, j)$ is greater than its two neighbors along the gradient direction $E_o(i, j)$. If so, keep $E_s(i, j)$ unchanged, otherwise, set it to 0.

5. Threshold the previous result by two different thresholds τ_1 and τ_2 (where $\tau_1 < \tau_2$) to obtain two binary images T_1 and T_2. Note that compared to T_1, T_2 has less noise and fewer false edges but larger gaps between edge segments.
6. Link edges segments in T_2 to form continuous edges. To do so, trace each segment in T_2 to its end and then search its neighbors in T_1 to find any edge segment in T_1 to bridge the gap until reaching another edge segment in T_2.

The result of applying Canny edge detector on a sample IR stereo pair is shown in Figure 3.3.c.

3.2.4 Kanade-Lucas-Tomasi (KLT)

In a very different approach from those mentioned above, Tomasi et al. [85] proposed a technique, known as KLT, which is not only capable of extracting local features in an image, but it also can track the extracted features in other views of (almost) the same scene. In this method, the displacement $d = [d_x \quad d_y]^T$ is computed to minimize the sum of the squared differences between consecutive image frames I and J (left and right images in our experiments) [68]:

$$\iint_W \left[I(x - \frac{d}{2}) - J(x + \frac{d}{2}) \right]^2 dx \ , \tag{3.10}$$

where W is a window of pixels around the considered feature point and $x = [x \quad y]^T$ is a pixel in the image. This nonlinear error can be minimized by solving its linearized version iteratively, through Taylor series expansion:

$$Zd = e \, , \tag{3.11}$$

where

$$\begin{aligned} Z &= \sum_{x \in W} g(x) g^T(x), \\ e &= \sum_{x \in W} g(x) \left[I(x) - J(x) \right] \end{aligned} \ , \tag{3.12}$$

and $g(x)=\frac{1}{2}\partial\left[I(x)+J(x)\right]/\partial x$ is the spatial gradient of the average image. Finally, features are selected as those points in the image for which both eigenvalues of Z are greater than a minimum threshold. Figure 3.3.d shows the performance of KLT on a pair of IR images.

3.2.5 Feature extraction used in our system (Phase Congruency)

Although the feature detectors reviewed in the previous section work very well for visible images, their performances on IR images are not reasonable enough (results of performing different feature detectors on some sample IR stereo images are illustrated in Figure 3.2. More detailed and quantitative comparative results are presented in chapter 5); either the number of detected features is not adequate, or the feature detectors are not robust enough to detect the same scene elements in the left and right images (repeatability characteristic [73]).

In our system, features are extracted from the phase congruency model. In comparison with the other studied approaches, the phase congruency model appears to provide a reasonable amount of stable and tractable features. In the following we briefly introduce the phase congruency model and its application on feature detection.

The Phase Congruency (PC) is a feature detection method which is based on the *Local Energy Model*, presented by Morrone et al. [61]. In contrast to the traditional approaches assuming that image features are located at points with maximal intensity gradient, the *Local Energy Model* makes the assumption that image features at frequency domain are located at points where the Fourier components are maximally in phase (see Figure 3.4). This is a more general definition than the one used by the traditional edge detection methods. Using the *Local Energy Model,* it is not required to make any assumptions about the shapes of the features being detected. This is an advantage over the other feature extraction techniques, which have to make assumptions about the shape of the feature to be detected. In addition, this method provides a degree of insensitivity to variation in illumination and contrast. For each point of the image in the phase congruency method, the congruency of phases at different frequencies (scales) is measured. Points with high phase congruency will get a high score, whereas points where the congruency is low get a low score. A PC feature edge map can then be computed by applying non-maxima suppression and thresholding.

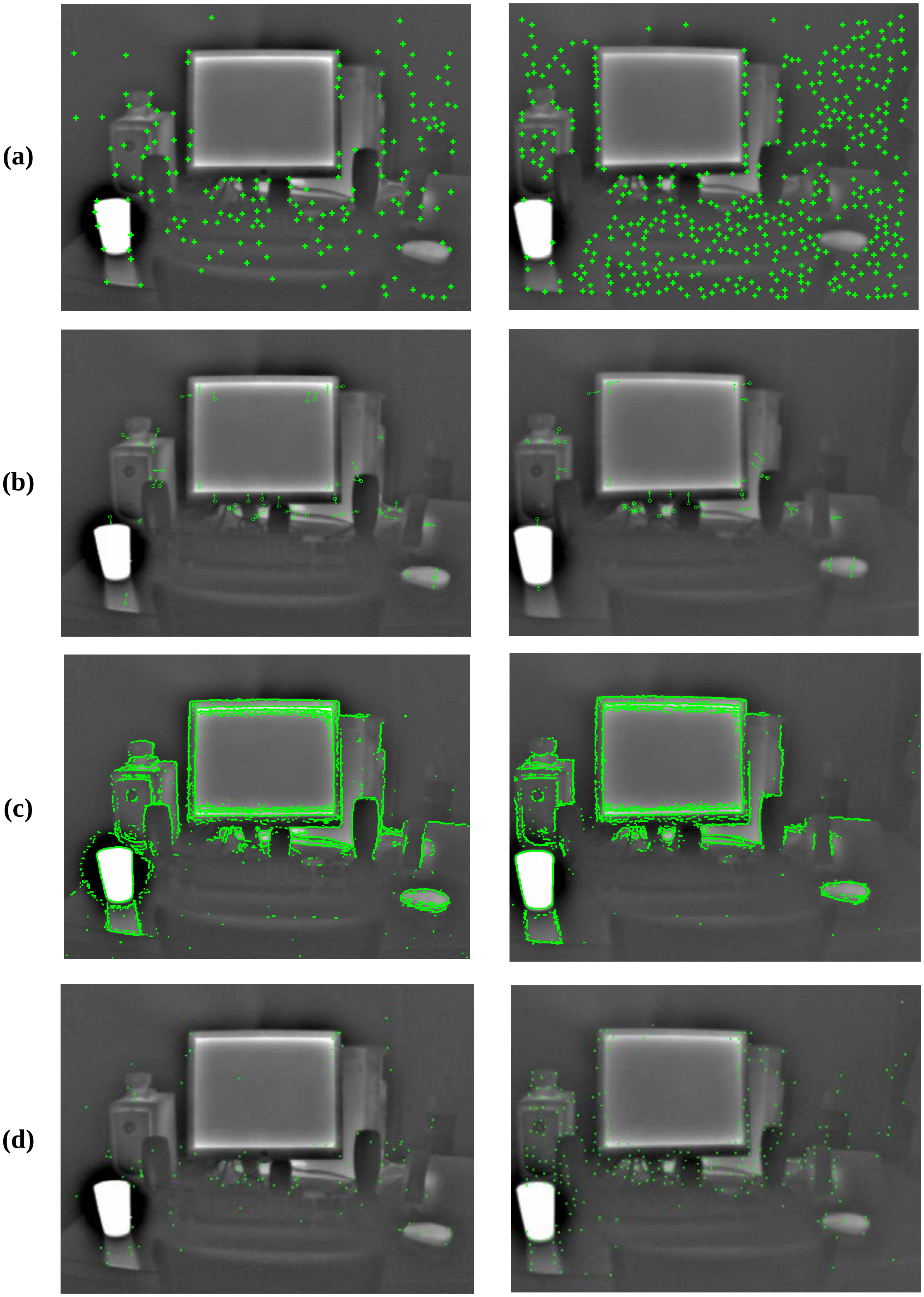
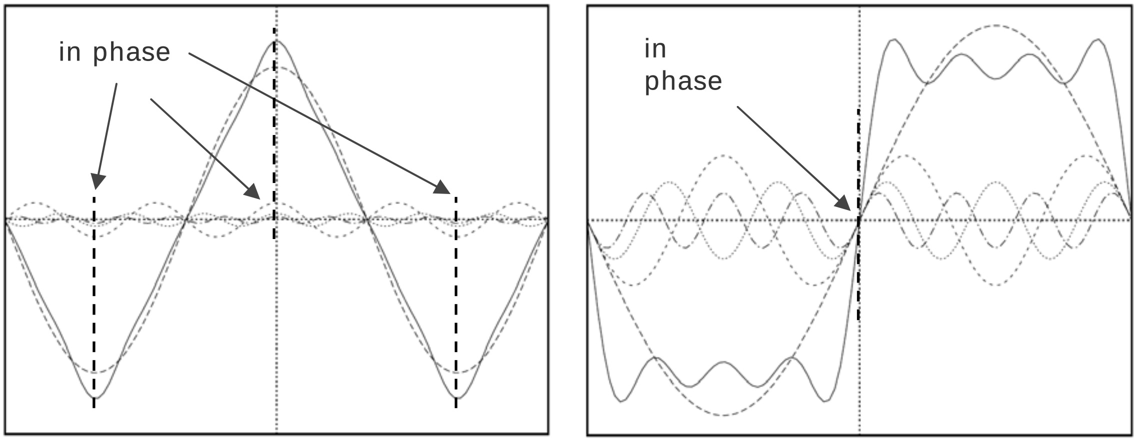


Figure 3.3: Performance of some state-of-the-art feature extractors on a sample IR stereo pair. (a), (b), (c) and (d) show the performances of Harris, DOG, Canny and KLT feature detectors.



(Square wave) (Triangular wave)

Figure 3.4: In both diagrams, the broken lines illustrate the first few terms of the respective Fourier series and the solid lines show the sum of these terms. At the point of the step at angles 0º and 180º in the square wave, all the Fourier components are in phase. Similarly, this is the case at peak and troughs of triangular wave, at angles 90º and 270º; (Reproduced from [49]).

A good way to describe and understand the concept of phase congruency model, is to consider the Fourier expansion of a 1D slice through an image [48]:

$$\begin{aligned} F(x) &= \sum_n A_n \cos(n\omega x + \phi_{n_0}) \\ &= \sum_n A_n \cos(\phi_n(x)) . \end{aligned} \tag{3.13}$$

In the above equation, A_n represents the amplitude of the n^{th} cosine component, ω is a constant (e.g., 2π), and ϕ_{n_0} represents the phase offset of the n^{th} component. The local phase of the Fourier component at position x is represented by the function $\phi_n(x)$. Using this function, Phase Congruency is defined as [61], [49]:

$$PC(x) = \max_{\bar{\phi}(x\in[0,2\pi])} \frac{\sum_n A_n \cos(\phi_n(x) - \bar{\phi}(x))}{\sum_n A_n(x)} . \tag{3.14}$$

The above equation is maximized when the local phase, $\phi_n(x)$, is equal to the amplitude weighted mean local phase angle of all Fourier terms, $\bar{\phi}(x)$, at the considered point (see Figure 3.5). Phase congruency as defined in equation (3.14) is relatively expensive to

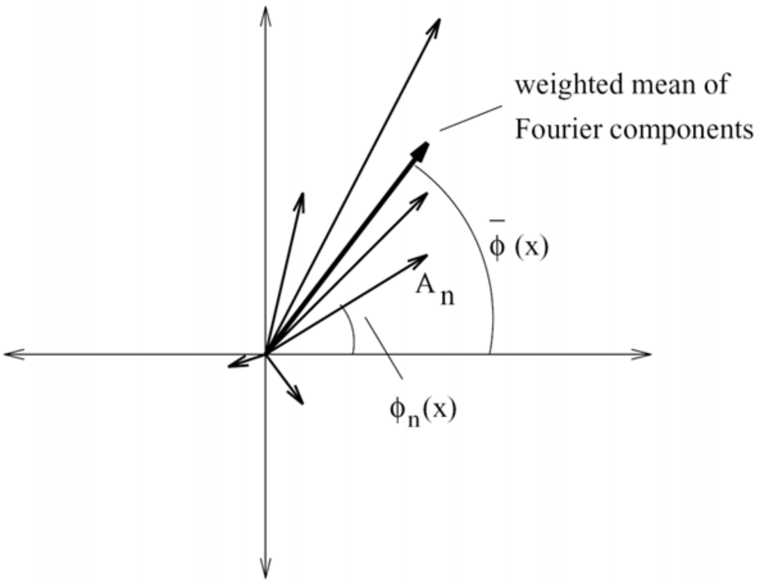


Figure 3.5: For a point in a signal, the polar diagram of the Fourier components is shown. A_n and ϕ_n represent the length and the local phase angle of each Fourier component, respectively; (Reproduced from [48]).

calculate. Venkatesh and Owens [91] have shown that the Phase Congruency function can be equivalently calculated by using the local energy function, which is mainly used in modeling iological vision. The local energy function for a 1-D signal is defined as [49]:

$$E(x) = \sqrt{F^2(x) + H^2(x)}\ , \tag{3.15}$$

where $F(x)$ is the signal with the DC component removed, and $H(x)$ is the Hilbert transform of $F(x)$ (a 90 deg. phase shift of$F(x)$). It was shown by Venkatesh and Owens that local energy is equal to phase congruency scaled by the sum of the Fourier amplitudes; that is,

$$E(x) = PC(x)\sum_n A_n\ . \tag{3.16}$$

From the above equation, it is obvious that the local energy function is directly proportional to the phase congruency function, and therefore, peaks in phase congruency will correspond to peaks in local energy.

Although using the local energy function for finding peaks in phase congruency is computationally more convenient, no dimensionless measure of feature significance can be provided, as the local energy function is weighted by the sum of the Fourier component amplitudes, which have units lumens [49]. This yielded the use of wavelet transform to obtain spatially localized frequency information in images for calculating phase congruency, which was first proposed by Morlet et al. [60]. In their approach, linear-phased filters were used, which are non-orthogonal wavelets based on complex valued Gabor kernels; each Gabor kernel consists of a cosine or even wave (forming the real part of the value) and sine or odd wave (forming the imaginary part of the value). Using even and odd filters in quadrature, the amplitude and phase of the signal can be computed for any particular frequency and spatial location.

Given I, the signal, and M_n^e and M_n^o, the even-symmetric (cosine) and odd-symmetric (sine) wavelets at a scale n, one can define the response vector of each quadrature pair of filters as:

$$[e_n(x), o_n(x)] = [I(x) * M_n^e, I(x) * M_n^o], \tag{3.17}$$

The amplitude of the transform at a given wavelet scale is given by:

$$A_n(x) = \sqrt{e_n(x)^2 + o_n(x)^2}, \tag{3.18}$$

and the phase is calculated by:

$$\phi_n(x) = a\tan 2(e_n(x), o_n(x)). \tag{3.19}$$

At each point x in a signal, a set of response vectors is obtained, one for each scale of a filter. These response vectors are used to produce the localized representation of the signal, and can be used in exactly the same manner as Fourier components are used to calculate the phase congruency.

Kovesi in [49], [50] follows the approach of Morlet et al. [60], but, rather than using Gabor filters, he uses Logarithmic Gabor functions that was suggested by Field [23]. Log-Gabor filters have a Gaussian transfer function when viewed on the linear and logarithmic frequency scale. They still maintain a zero DC component in the even-symmetric filter, while allowing arbitrarily large bandwidth filters to be constructed.[3]

The 2D Log-Gabor filter is constructed in the frequency domain. In the spatial domain, it can only be numerically constructed using the inverse Fourier transform. Log-Gabor filters

[3] A zero DC value cannot be maintained in Gabor functions for bandwidths over one octave [49].

consist of two components, namely the radial filter component and the angular filter component. The radial filter has the transfer function [49]:

$$G(\omega) = \exp\left(\frac{-(\log(\omega/\omega_0))^2}{2(\log(k/\omega_0))^2} \right), \tag{3.20}$$

where ω_0 is the center frequency of the filter and k determines the bandwidth of the filter in the radial direction. The angular component, which controls the orientation that the filter responds to, has the Gaussian transfer function [49]:

$$G(\theta) = \exp\left(\frac{-(\theta-\theta_0)^2}{2T(\Delta\theta)^2} \right), \tag{3.21}$$

where θ_0 is the orientation angle of the filter, T is a scaling factor, and $\Delta\theta$ represents the orientation spacing between the filters.

The Log-Gabor filters are produced by multiplying the radial and angular components together (see Figure 3.6). A bank of 24 Log-Gabor filters at 4 frequencies and 6 orientations is used in our implementation.

Kovesi [49], [50] developed a modified measure of phase congruency (via wavelets), consisting of the cosine minus the magnitude of the sine of the phase deviation, which produces a more localized response (than equation (3.14)). This new measure also incorporates noise compensation:

$$PC_2(x) = \frac{\sum_n W(x) \lfloor A_n(x)\Delta\Phi(x) - T \rfloor}{\sum_n A_n(x) + \varepsilon}. \tag{3.22}$$

where $\Delta\Phi(x) = (\cos(\Phi_n(x) - \overline{\Phi}(x)) - \left|\sin(\Phi_n(x) - \overline{\Phi}(x))\right|)$. In the above equations, $\Delta\Phi(x)$ is the phase deviation, and $W(x)$ is a factor that weights for frequency spread. ε is a small constant that is used to avoid division by zero. Note that only energy values that exceed T, the estimated noise influence, are regarded in the result; and the $\lfloor\ \rfloor$ denote that the enclosed quantity is equal to itself when its value is positive, and zero otherwise.

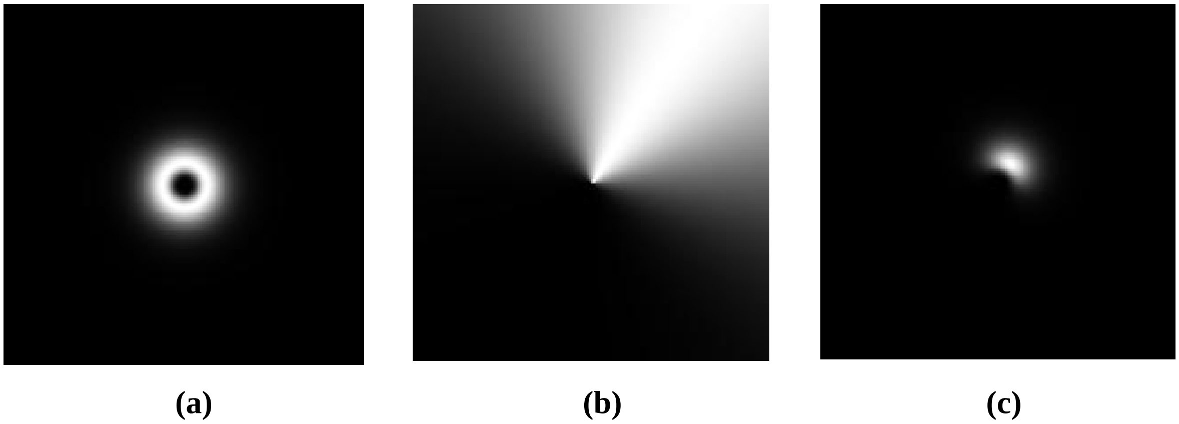


Figure 3.6: (a) Radial log-Gabor component of the filter, (b) angular component of the filter, (c) product of (a) and (b) to produce the frequency domain representation of the log-Gabor filter.

By combining phase information at different orientations into a covariance matrix, and calculating the minimum and maximum moments, a highly localized representation is produced that can be used to detect both edges and corners in a contrast invariant manner. In order to extract edge features from the phase congruency model, the following measures are computed at each point in the image [50]:

$$a = \sum (PC(\theta)\cos(\theta))^2 , \tag{3.23}$$

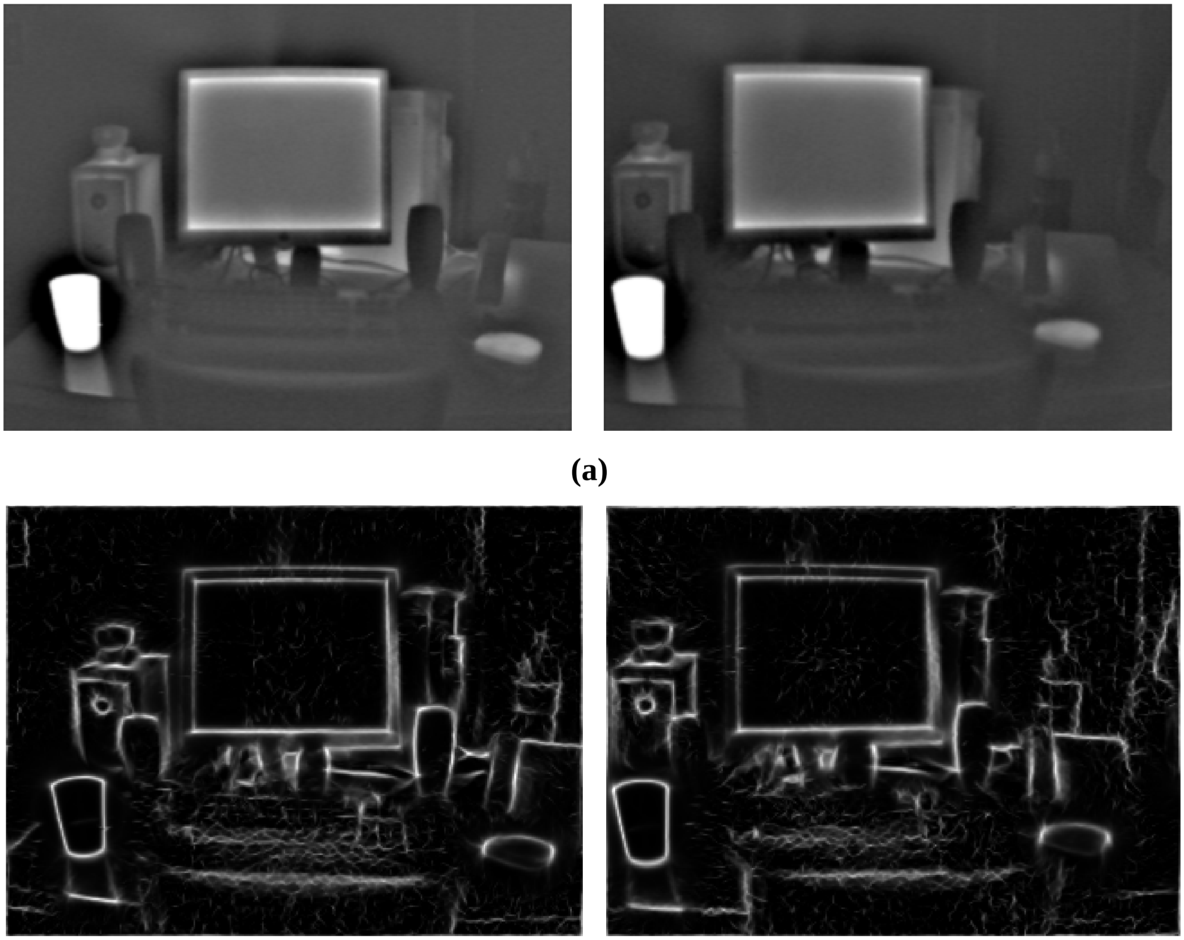
$$b = 2\sum (PC(\theta)\cos(\theta))(PC(\theta)\sin(\theta)) , \tag{3.24}$$

$$c = \sum (PC(\theta)\sin(\theta))^2 , \tag{3.25}$$

where $PC(\theta)$ represents the phase congruency value determined at orientation θ, and the sum is performed over all discrete orientations used. Now, the edge coefficient of each pixel is computed using the maximum moment of phase congruency covariance:

$$M = \frac{1}{2}(c + a + \sqrt{b^2 + (a-c)^2}) . \tag{3.26}$$

PC edge maps extracted from a sample IR stereo are illustrated in Figure 3.7. The edge maps are further processed by a locally adaptive thresholding strategy in order to binarize the results. Furthermore, the following two-step process is performed on the binary map in order to remove the extra unnecessary points: (i) removing the connected regions whose areas are smaller than a



(a)

(b)

Figure 3.7: (a) original IR stereo images (b) the phase congruency edge maps.

given threshold and are not reliable, (ii) thinning the connected regions using a set of successive morphological operations.

3.3 Feature matching

Once the reliable and tractable features are extracted from the left and right images, the next question is how these extracted features should be described to make the comparison and matching possible. There are a large variety of possible descriptors and associated distance metrics which emphasize different image properties like pixel intensities, color, gradient magnitude and textures. However, the selection of a reasonable feature descriptor depends basically on the type of the extracted features and the underlying extraction technique. For example, for the regions extracted by the Difference-of-Gaussians technique, Lowe [56] proposed a scale and rotation invariant description method, which is based on the distribution of the gradient magnitudes and orientations, computed during the extraction of the regions.

In our system, features are described by a set of Log-Gabor coefficients at different scales (e.g. 4) and orientations (e.g. 6), computed during the extraction of feature points by the phase congruency model. Similar to Gabor wavelets, Log-Gabor wavelets are biologically motivated convolution kernels restricted by a Gaussian function. Gabor and Log-Gabor wavelet coefficients have been used as descriptors in several applications including object recognition [51], face recognition [93] and content based image retrieval [26]. The Log-Gabor wavelet kernels used in our system are illustrated in Figure 3.8.

Once the features are described, the matching between the right and the left images is possible through a simple and straightforward algorithm: each feature at (x, y) in the first image (left/right) is compared with a set of potential features in the other image (right/left), located within a $w_x \times w_y$ rectangular window, W, centered on (x, y), where w_x and w_y are the maximum expected disparities in horizontal and vertical axes, respectively. Features are compared using the cosine similarity function, which has also been used in [51] and [93] for comparing wavelet-based description vectors:

$$S(F, F') = \frac{\sum_j f_j f'_j}{\sqrt{\sum_j f_j^2 \sum_j f'^2_j}}, \tag{3.27}$$

where F and F' are two sets of Log-Gabor coefficients' magnitudes to be compared, and f_j and f'_j are the j^{th} coefficients in F and F', respectively. For each feature in the first image, the feature in the second image which maximizes the similarity measure is then selected as the tentative corresponding point.

The performance of any correspondence method is impaired by 'occlusions' (points with no counterpart in the other image) and 'spurious matches' (false corresponding pairs created by noise). Appropriate constraints reduce the effects of both phenomena. We impose two such constraints on our system: namely, the *left-right consistency constraint*, and the *uniqueness constraint*. The *uniqueness constraint* means that a given feature from one image can be

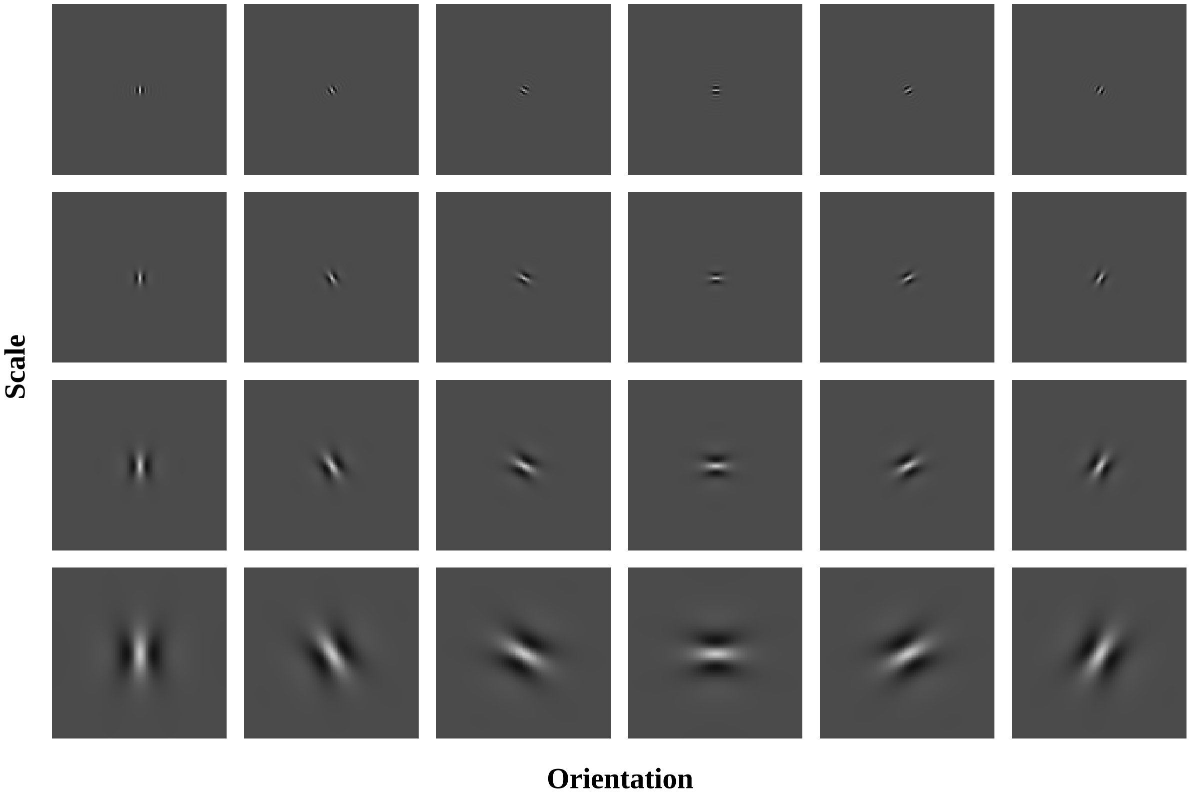


Figure 3.8: Log-Gabor filters at four scales and six orientations. Only the real parts of the filters are shown.

matched to only one feature from the other image [58]. On the other hand, the *left-right consistency* constraint is on the basis of checking that a valid match point be equally matched and have the same disparity in both left-right and right-left directions. Fua in [27] proposed a technique for left-right consistency, which is illustrated in Figure 3.9.

3.4 Matching refinement

In general, since for a given point in one stereo image, a corresponding point in the other image may not exist, due to occlusion or missing parts, the feature matching process is usually considered as an ill-posed problem. Furthermore, if the feature points are not located in a sufficiently textured area (which is common in IR images), the matching algorithm may fail to correctly find the correspondences, thus, producing outliers.

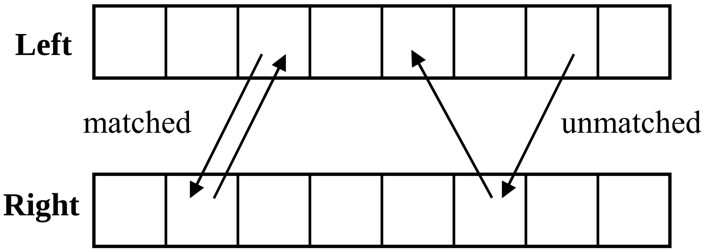


Figure 3.9: Left-right consistency checking. Each arrow represents the matched elements and the matching direction (left-right or right-left). A match between two elements is valid if they are matched to each other in both directions.

In this section we describe a method for removing the outliers, using the fundamental matrix for estimating the epipolar geometry and RANSAC strategy to robustly fit the Fundamental Matrix to the matched points.

3.4.1 Epipolar geometry

The intrinsic projective geometry between two views is called the geometry of stereo, also known as *epipolar geometry* [42], which is independent of scene structure, and only depends on the cameras' internal parameters and relative pose. First of all, we need to establish some basic notations.

Notation. Let's consider a scene point M that is visible in both stereo cameras. Given its projection, m_l, on the image plane formed by the left camera, its corresponding projection, m_r, on the image plane formed by the right camera, has to be located on the *epipolar line*, l_{mr} (*epipolar constraint*). The epipolar line is the intersection between the *epipolar plane*, Π, (defined by the optical centers of the stereo cameras C_l and C_r, and the scene point under investigation M) and the image planes, π_l and π_r [29]. Figure 3.10 shows the epipolar geometry. As can be seen, the epipolar lines, l_{ml} and l_{mr}, show the relative orientation and position of the scene points in the stereo images. The points e_l and e_r in Figure 3.10 are the *epipols* and are defined by the intersection of the image planes with the baseline $\{C_l, C_r\}$.

The most common way to represent the epipolar geometry is through a 3×3 matrix called the *Fundamental Matrix* [42], noted as F. The fundamental matrix can be derived from the

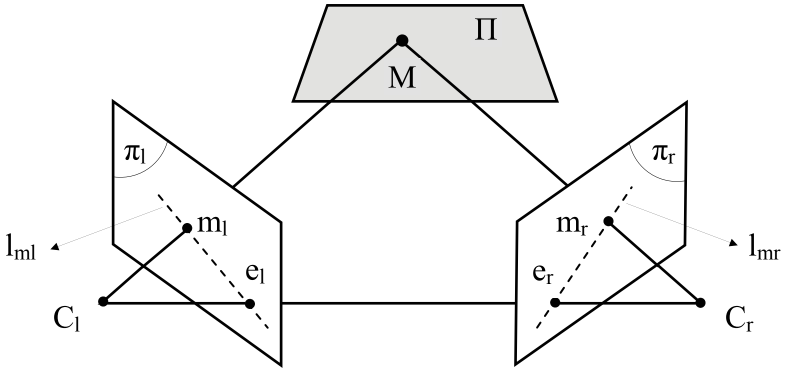


Figure 3.10: The epipolar geometry

mapping between a point and its epipolar line: for a given point m in the first image, the projective representation of the epipolar line in the second image l'_m, is given by:

$$l'_m = Fm\,. \tag{3.28}$$

Since the point m' corresponding to m belongs to the line l'_m by definition, therefore: $m'^T l' = m'^T Fm = 0$. So, the fundamental matrix satisfies the condition that for any pair of corresponding feature points $m \leftrightarrow m'$ in the two images:

$$m'^T Fm = 0\,, \tag{3.29}$$

where $m = (x, y, 1)^T$ and $m' = (x', y', 1)^T$, in homogeneous coordinates.

The simplest method of computing the fundamental matrix is the 8-point algorithm, presented originally by Longuet-Higgins [55]: Given a sufficient number of matched points $m_i \leftrightarrow m'_i$, (at least 8), the equation (3.29) can be used to compute the unknown matrix F. In particular, writing $m = (x, y, 1)^T$ and $m' = (x', y', 1)^T$, each matched pair gives rise to one linear equation in the unknown entries of F. The coefficients of this equation are easily written in terms of the known coordinates m and m'. Specifically, the equation corresponding to a pair of $(x, y, 1)$ and $(x', y', 1)$ is:

$$[xx' \quad yx' \quad x' \quad xy' \quad yy' \quad y' \quad x \quad y \quad 1]\, f = 0\,, \tag{3.30}$$

where $f = [F_{11} F_{12} F_{13} F_{21} F_{22} F_{23} F_{31} F_{32} F_{33}]^T$ is a 9-vector containing the elements of the fundamental matrix F. By stacking eight of these equations in a matrix A, we can obtain a linear system of the form $Af = 0$. The least-squares solution for f is the singular vector corresponding to the smallest singular value of A, that is, the last column of V in the $SVD(A) = UDV^T$. The solution vector f found in this way minimizes $\|Af\|$ subject to the condition $\|f\| = 1$.

In order to reduce the numerical complexities while calculating the matrix F, Hartley in [41] proposed the normalized 8-point algorithm. In this approach, the center of corresponding points is translated to the origin of the image reference frame and then the corresponding points are scaled so that the average distance from the origin becomes equal to $\sqrt{2}$. Finally, after the calculation of matrix $\hat{F}$ by using the 8-point algorithm, it is converted to the matrix F of corresponding points before normalization using: $F = T_r^T \hat{F} T_l$, where T_r and T_l are the transformation (normalization) matrices for the right and left images, respectively.

3.4.2 Fundamental matrix RANSAC fit

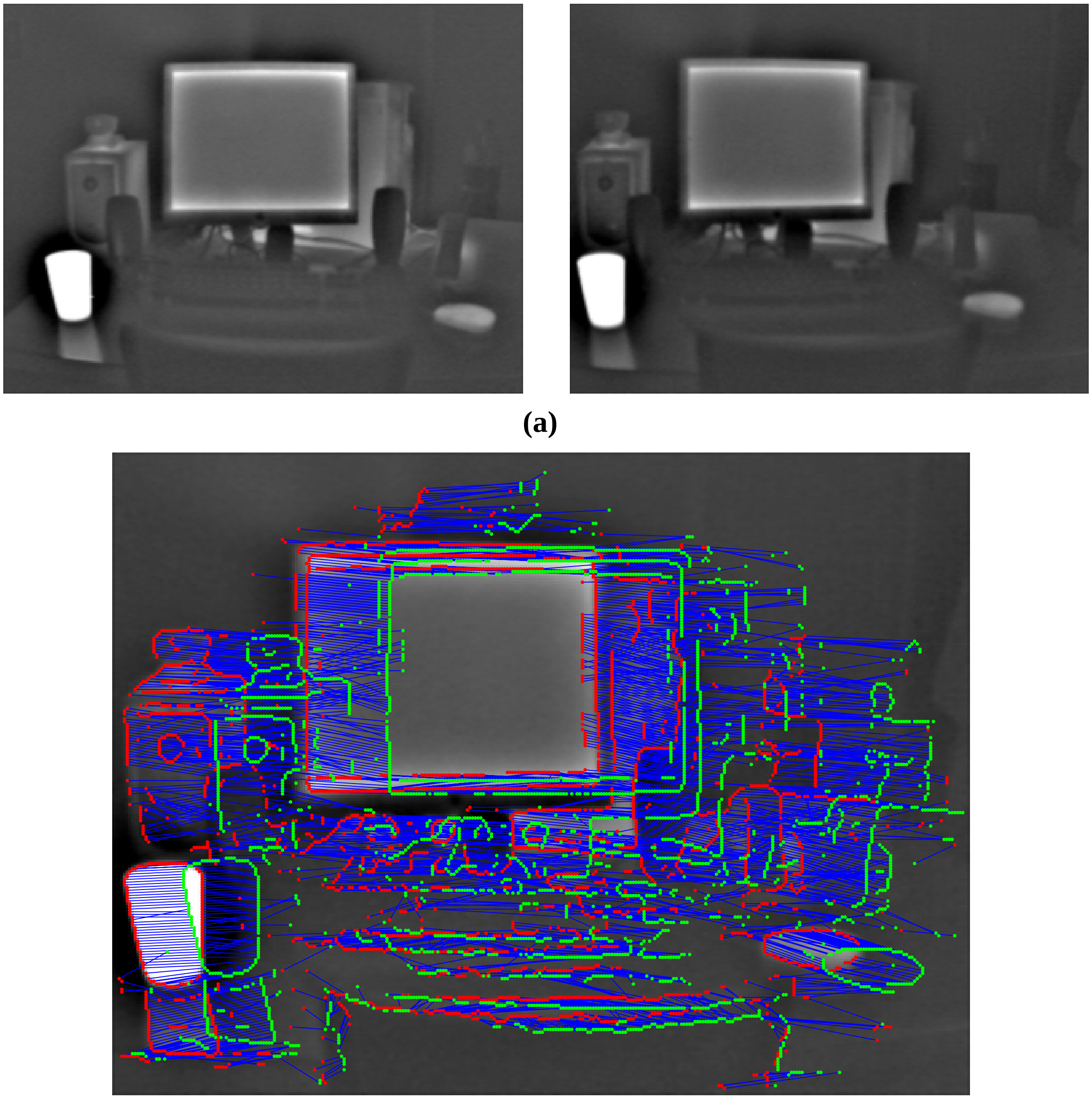
The RANSAC (or RANdom SAmple Consensus) algorithm [24], [86] is an algorithm for robust fitting of models in the presence of many data “outliers”. We use this algorithm to robustly fit a fundamental matrix to a set of putatively matched image points and obtain a subset (called “inliers”), that is consistent with the epipolar geometry.

Given N matched feature pairs, the RANSAC algorithm iteratively performs the following steps: (i) selecting p sample pairs (p =8, the minimum number required to compute a fundamental matrix) at random, (ii) computing the fundamental matrix (as described in Section 3.4.1), and (iii) determining to what degree all the available matches support the epipolar constraint. The random sampling is repeated m times until at least one random sample contains only good matches with probability $P = 1-(1-(1-\varepsilon)^p)^m$, ($\varepsilon$ is the percentage of outliers). After fitting the fundamental matrix, any matched points which do not satisfy the fitted model, are returned as outliers.

In our approach, once the matched points are extracted and the left-right consistency is applied (using the algorithm described in Section 3.3), we fit a fundamental matrix to the remained matched points using the RANSAC technique (as described above) to detect and remove the outliers. Figures 3.11 and 3.12 illustrate the matching performance, the recovery of the epipolar lines and the associated disparity map for one sample stereo pair of our data set. The blue lines indicate the features matched in both views, and the red and green circles represent the feature points in the right and left images, respectively. It can be observed that the incorrectly detected epipolar lines derived from the mismatched features are successfully identified and removed.

3.5 Chapter summary

In this chapter, we proposed a method for feature-based matching to produce sparse depth maps that can be used to generate semi-dense depth fields. Our method was composed of three main steps, namely, feature extraction, feature matching, and matching refinement. In the feature extraction step, we studied the performances of some of the state-of-the-art feature detectors and compared them with the performance of the phase congruency model, used in our system for feature extraction. Analyzing the results qualitatively, phase congruency method achieved the best performance on the texture-less and low-resolution IR images (quantitative analysis of the performance of different detectors on IR images are presented in Chapter 5). In the feature matching step, a set of Log-Gabor wavelet coefficients in different orientations and frequencies (computed during the calculation of phase congruency) were used to describe the extracted features, and a cosine similarity function was employed to match the features in left and right images. Epipolar geometry and left-right consistencies were then used in the third step in order to detect and eliminate the outliers. More experiments with our methodology and quantitative analysis of the results will be presented in Chapter 5.



(a)

(b)

Figure 3.11: (a) shows the original IR stereo pair; (b) displays the right image overlaid with all matched feature points from the right and left images, along with the correspondences (with blue lines);

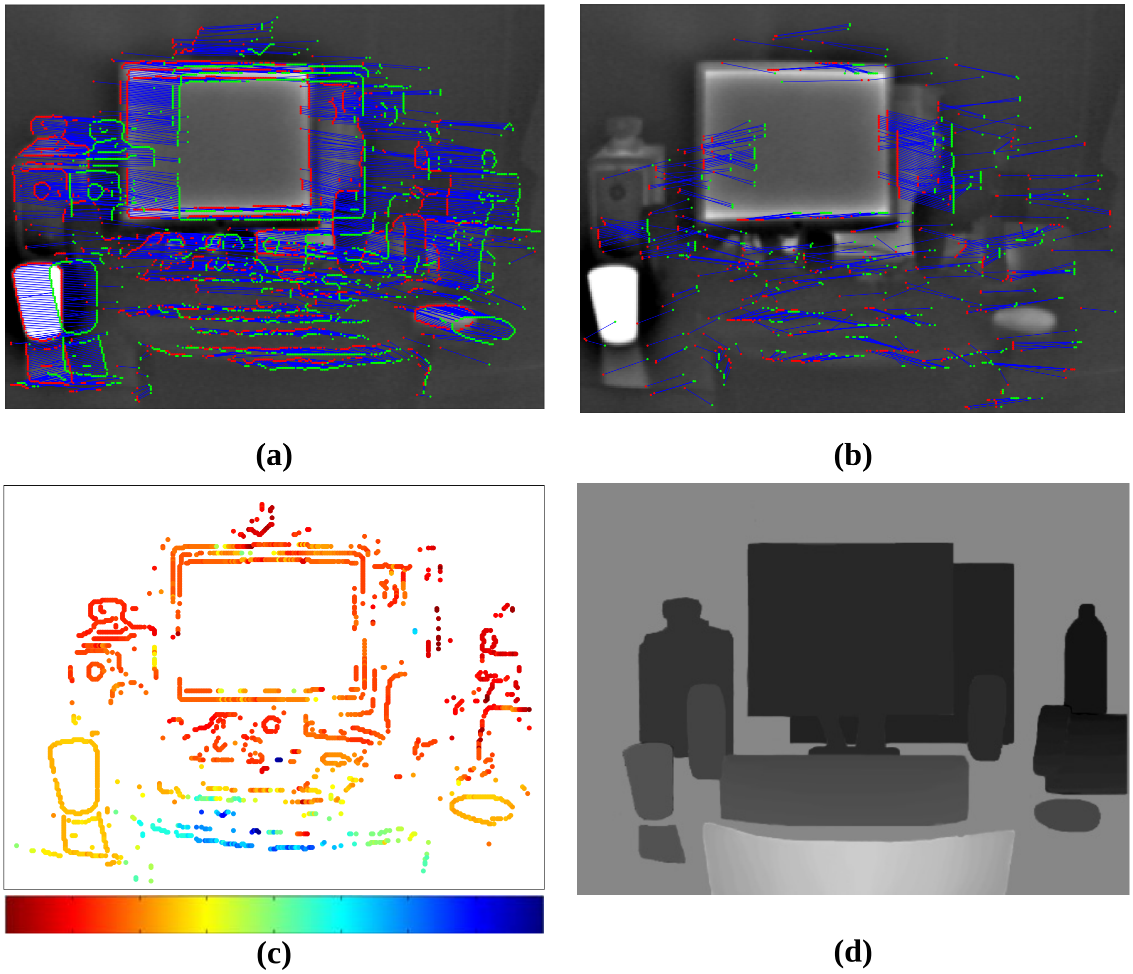


Figure 3.12: (a) displays Figure 3.11.b after outlier elimination; (b) displays the detected outliers. (c) Sparse disparity map generated from an IR stereo pair. Darkest blue, in disparity map, indicates objects closest to the camera, and darkest red indicates objects farthest from the camera; (d) the ground truth data that was obtained with respect to the depth ordering of the presented objects in the scene, to make the evaluation of the resulted disparity map possible. Closer objects are shown with brighter gray values, while farther objects are shown with darker gray levels.

Chapter 4

Surface reconstruction

Once the sparse depth map is generated, the next step is to reconstruct the surface in order to find more detailed depth information. Surface reconstruction is the process of reconstructing a piecewise smooth surface from a set of noisy measurements (known data). Besides the smoothness, in many applications the reconstructed surface is also required to identify and preserve the discontinuities.

In the feature-based stereo imaging, measurements are often obtained through the feature matching between the left and the right images (i.e., depth constraints). Therefore, usually a very sparse and irregular sampling is obtained which very likely doesn't contain measurements in some parts of the image (e.g., where no feature is detected or matched in the left and right images). This makes the reconstruction problem ill-posed and therefore some additional constraints are required to make it well-posed[1].

In this chapter, we first present a brief study on some surface fitting methods. Then we describe the problem of identifying and preserving discontinuities, and finally we present different steps of our surface reconstruction technique.

[1] A common approach to solve this ill-posed problem is through the regularization technique originally proposed by Tikhonov [84], which restricts the solution to be a smooth function. See section 4.3.2 for more details.

4.1 Related work

Most of the feature-based stereo matching methods [34], [58] incorporate a surface reconstruction (interpolation) technique to densify their result. Generally, the techniques used for interpolation or approximation of possibly sparse depth maps (produced by a feature-based stereo matching) by a surface, can be categorized in at least two forms: (i) minimization of spline[2] functional and (ii) directly fitting polynomial based surface patches. These categories are briefly reviewed in the following:

Minimization of spline functional. Influential works in this category are those of Terzopolous [83], Grimson [34] and Blake and Zisserman [9]. Grimson proposed a visual surface interpolation technique in which depth information is obtained from stereo images. More specifically, this approach deals with fitting the best surface to a sparse set of depth values produced by the Marr and Poggio [58] stereo algorithm. For smooth interpolation of visual surfaces from depth measurements, Grimson minimized the quadratic variation E of the surface $f(x,y)$:

$$E = \iint_R f_{xx}^2 + 2f_{xy}^2 + f_{yy}^2 \quad dxdy \,. \tag{4.1}$$

In the above equation, R is the image region in which the depth constraints are specified. This functional is equivalent to the energy of a thin plate, whose minimization produces a function called *thin-plate splines*. Grimson developed an iterative algorithm based upon the biharmonic equation that results from applying Euler's equations to minimize E [10]. Since this approach did not deal with surface or orientation discontinuities, it cannot be applied to the general surface reconstruction problem.

A significant disadvantage of the thin-plate spline is that, its global C^1 continuity gives rise to undesirable exceeded values near large gradients. To overcome this deficiency of global smoothness constraints Terzopoulos [82], [83] introduced a general class of controlled-continuity stabilizers, which provides the necessary control over smoothness. This non-quadratic stabilizing functional is a 2D extension of the spline under tension functional and is

[2] A spline funstion for interpolation is considered as a minimizer of suitable measures of smoothness subject to some interpolation constraints.

recognized as a weighted convex combination of the thin-plate and membrane generalized spline kernels:

$$\varepsilon(f) = \iint_{\Omega} \rho(x,y)\Big\{\tau(x,y)(f_{xx}^2 + 2f_{xy}^2 + f_{yy}^2) \\ + [1-\tau(x,y)](f_x^2 + f_y^2)\Big\}\ dxdy \quad , \tag{4.2}$$

where $\Omega \subset \Re^2$ refers to the images domain and $f(x,y)$ is the approximation function. The 2D functions $\rho(x,y)$ and $\tau(x,y)$, also called continuity control functions, involve an explicit representation of depth and orientation discontinuities, respectively.

The stabilizer is controlled as follows: when $\rho(x,y) = \tau(x,y) = 1$, which is the case over smooth surface patches and away from depth and orientation discontinuities, the solutions are *thin-plate splines*. At discontinuities in surface orientation, where $\rho(x,y) = 1$ and $\tau(x,y) = 0$, the solutions are the *membrane splines*. Finally, near depth discontinuities, $\rho(x,y) = 0$, the solution is constrained to agree with the measured data as closely as possible. Setting the continuity control functions ρ and τ requires a prior knowledge of the location of the discontinuities. Terzopolous in [83] proposed two ways to identify the discontinuities, namely local validation by bending moment method and the variational continuity control.

Directly fitting polynomial based surface patches. In this form of surface fitting, the assumption on disparity smoothness is implemented by fitting planar or quadratic surface patches locally to the measurements (e.g., Hoff and Ahuja [44], Eastman and Waxman [21], Faugeras et al. [22]).

Hoff and Ahuja in [44] presented an approach that integrates the processes of feature matching, contour detection, and surface interpolation, into a single process. The integration is implemented in a multiresolution hierarchy of surface maps. In their algorithm, a Laplacian of Gaussian operator is used to extract edgels and then matching is performed in both left-to-right and right-to-left passes. The possible surfaces (first planar and then quadratic) are fitted to the points within a circular patch around each depth point across the image, using Hough transform. The objects in the scene grow smaller planar patches at known surface points until they merge or break at discontinuities (which are detected by fitting a bipartite patch at various orientations). Finally, a piecewise smooth surface evolves. This approach achieves very promising results; however it fails when the matching features are too sparse. Also, the surface discontinuities usually are not located accurately.

Eastman and Waxman [21], also use fitting planar patches to a set of offset values. In their approach, those matching candidates, which have a disparity gradient greater than 1, or deviate the residual of the fit more than three times, are identified as outliers and therefore are eliminated. A new fit is made and the process is iterated up for to three times. Finally, the disparity offset with the minimum residual is accepted.

4.2 Discontinuities: detection and preservation

Surface reconstruction involves not only interpolating the smooth surfaces over uniform regions but also locating and preserving the *discontinuities* that bound these regions, since very often they carry the most important information in the scene [89]. Standard regularization techniques (e.g., [34], [67]) impose a smoothness criterion over the whole image and therefore are not capable of preserving the discontinuities in depth. More advanced techniques which consider the discontinuity detection and preservation, can be generally categorized into three groups: (i) methods which detect discontinuities prior to reconstruction (e.g., in a work by Grimson and Pavlidis [35], discontinuities are detected by monitoring local statistics of the residuals of a local planar approximation to the depth data); (ii) those methods which integrate the surface depth discontinuity information into the surface reconstruction problem (e.g., Blake suggested in [8] the idea of including a discontinuity penalty term as part of the minimal energy formulation for piecewise continuous reconstruction); and (iii) methods which detect discontinuities subsequently (e.g., Terzopoulos in [83] applies a post processing local validation on the reconstructed surface to locate and recover the discontinuities).

4.3 Our technique

Since in our application (depth from IR stereo), the prior knowledge about the location of discontinuities can always be generated from one of the stereo images (referred to as reference image), we chose to handle the problem of depth discontinuity detection and preservation by first extracting the edge map of the reference image (since the edge features provide the exact, non-blurred locations for the discontinuities). Then we perform segmentation to obtain homogeneous regions, where the disparity can be assumed to vary smoothly inside each region. Finally, the surface in each region is reconstructed independently.

4.3.1 Segmentation

Segmentation, as the process of separating touching objects in an image, is a very difficult image processing task. Among the available techniques for image segmentation, we chose to use a watershed-based method [7], due to its simple and efficient framework, and because of its ability of producing more stable segmentation results for IR images.

In geography, a *watershed* is the ridge that divides areas drained by different river systems. In this sense, a *catchment basin* means a geographical area from which rainfall flows into a river or reservoir, and *watershed lines* are the dams which separate the basins [95]. The *watershed transform* applies these ideas to image processing to enable the solution of a variety of computer vision problems. Considering the input image as a topological surface, where the values of $f(x, y)$ are interpreted as heights, the watershed transform finds the catchment basins and ridge lines in such an input image. When the aim is to use watershed transformation for gray-scale image segmentation, the main task is to change the original gray-scale input image to an intermediate representation, whose catchment basins are the objects or regions we want to distinguish. Among the widely used intermediate representation techniques, employed together with watershed transformation, are the *distance transform* and *gradient*. Since we are interested in identifying regions in the reference image where the disparity/depth information varies smoothly (or in other word, regions with no major discontinuity involved), we chose to use the distance transform of the edge map, as the intermediate representation for the watershed segmentation of the reference IR image.

1	1	0	0	0
1	1	0	0	0
0	0	0	0	0
0	0	0	0	0
0	1	1	1	0

Binary Image

0.00	0.00	1.00	2.00	3.00
0.00	0.00	1.00	2.00	3.00
1.00	1.00	1.41	2.00	2.24
1.41	1.00	1.00	1.00	1.41
1.00	0.00	0.00	0.00	0.00

Distance Transform

Figure 4.1: Distance Transform (DT)

The distance transform of a binary image (i.e., edge map) is a simple concept [31]. It represents the Euclidean distance from each pixel with its value of 1 to the nearest nonzero-valued pixel as can be seen in Figure 4.1. The peaks of the distance transform are in the middle of the regions of interest. The idea is to run the watershed transformation using these peaks as the starting points (*markers*). For this, we invert the distance transform so that the peaks become the regional minima (*catchment basin*), and assign all zero pixels of the distance transform (those pixels which are located in edge points of the reference image) by $-\infty$. Figure 4.2.c shows the distance transform of the binary map in Figure 4.2.b. After proper preprocessing of the distance transform, the watershed is applied to obtain the image segments (Figure 4.2.d).

A well-known problem of watershed segmentation is a tendency towards *oversegmentation*[3]. Since the watershed generates a large number of segmented regions, the main challenge is to make a proper choice and select and merge the relevant regions only. There are several approaches for merging watershed regions to obtain larger and more interesting image segments, like [94] and [39]. We also applied some image pre-processing and post-processing, in order to suppress the oversegmentation and obtain a concise region representation:

Pre-processing. We preprocessed the distance transform with a simple 5×5 median filter. The aim of this filter is to get rid of spurious minima and improve the result of watershed segmentation. Larger size filters may cause over-smoothing that will weaken important edges, therefore, causing an incomplete image representation.

Post-processing. This stage consists of, first, merging some of the basins in a proper way by removing irrelevant watershed lines, and then merging those segmented regions that do not

[3] Oversegmentation occurs when every local minimum, even if insignificant, forms its own catchment basin.

satisfy the minimum valid region size criterion, with their neighbors.

In the first step, the irrelevant watershed lines, i.e., regions' common boundaries, are eliminated and adjacent regions are merged, if:

$$N_E < T_1, \qquad \frac{N_E}{N_L} \le T_2, \qquad N_L > T_3, \tag{4.3}$$

where, N_E is the number of edge pixels on the watershed line (common boundary), and N_L is the length of the watershed line. T_1, T_2 and T_3 are preset thresholds. In all our experiments T_1, T_2 and T_3 are set to 5, 1/12 and 5, respectively.

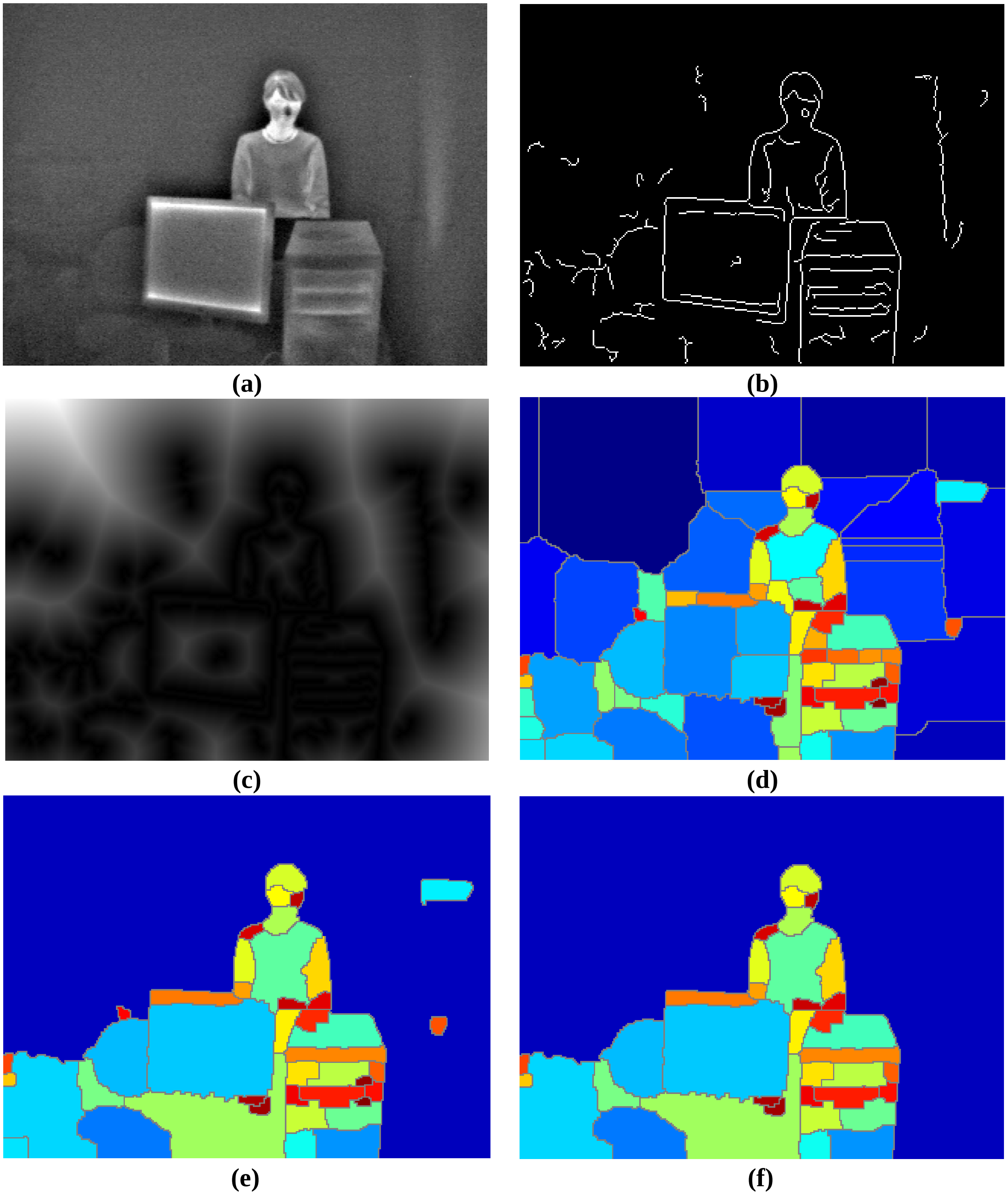
In the second step, we remove small size regions[4]. If a certain region has an area smaller than a threshold T_A, then its borders with the neighboring regions are searched, and this region is merged with the neighboring region which has the widest border. It is clear that threshold T_A is inherently application dependent, because the minimum object area can vary significantly for different applications.

As an illustration of the efficiency of our post-processing technique, the final segmentation of a sample IR image is displayed in Figure 4.2.f. In our experiments, we used $T_A = 50$ (for images of size 320×240), so that regions with an area smaller than 50 pixels were merged.

4.3.2 Interpolation of each segment

We assume that for the homogeneous regions obtained from segmentation, the disparity varies smoothly and depth discontinuities coincide with the boundaries of those regions ([75], [98]), which holds true for most natural images (including IR images). Therefore, to reconstruct the surface and preserve the discontinuities in the result, we reconstruct each region individually based on the available measurements within that region. In order to find out which surface fitting method is more robust to be employed in our final system, we perform a comparative study on different surface fitting techniques, using several synthetic data (10 images). In the following we first describe different surface fitting techniques and then present our comparative experiment.

[4] Usually, realistic objects in an image exist within a range of sizes; therefore, it is possible to impose a constraint on object area for segmented regions.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 4.2: (a) left image of an IR stereo pair (reference image); (b) binary edge map; (c) the distance transform of the binary edge map; (d) initial segmentation result; (e) segmentation result after removing irrelevant watershed lines; (f) segmentation result after removing irrelevant watershed lines and removing/merging small regions.

Surface fitting techniques. In general, there are two forms of expressing surfaces: implicit and explicit forms [10]. Explicit surface reconstruction techniques are commonly characterized as optimal approximation problems dealing with a class of generalized spline functions. The smoothness constraint is the basic component in all these formulations which acts as a stabilizing functional. These surface fitting techniques, usually lead to robust algorithms, due to their strong formulation.

The depth map obtained from a stereo correspondence can be interpreted in the explicit form $z = f(x, y)$, where z is the distance from the camera to an object point in a scene and (x, y) are the image projection coordinates of the considered object point. The depth function $f(x_k, y_k)$ can be computed for all image positions (x_k, y_k) by minimizing the error function:

$$E(f) = \sum_k (z_k - Cf(x_k, y_k))^2 + \lambda S(f). \tag{4.4}$$

In the above equation, the first term is the data constraint, i.e., the residual error in fitting the surface f to the known measurements, z_k. C is a matrix that maps the depth domain to the measurements domain. In the second term, $S(f)$ represents the smoothness requirement applied to f, and λ is a regularization constant which tune the tradeoff between the data constraint and the smoothness constraint. Two common choices for the functional $S(f)$ is the first-order form (membrane):

$$S(f) = \iint [(D_x f)^2 + (D_y f)^2] dxdy, \tag{4.5}$$

and the second-order form (thin-plate):

$$S(f) = \iint [(D_x^2 f)^2 + 2(D_x D_y f)^2 + (D_y^2 f)^2] dxdy, \tag{4.6}$$

where, D_x and D_y are the differential operators with respect to x and y, respectively. The Euler-Lagrange differential equation is used to solve these variational problems.

Unlike the explicit methods, the implicit surface reconstruction techniques often try to extract global properties (parameters) of a surface [10]. For reliable and accurate parameter estimation, all the points that are used for the estimation are required to lie on the same surface, the surface for which the parameters are estimated. The implicit form of the surface is expressed as the function: $f(x,y,z) = const.$, where (x,y,z) are the Cartesian coordinates of the surface points. The depth (z) of a point located at (x,y) according to the planar and quadratic models, is given by the following equations, respectively:

$$z = a_1 x + b_1 y + c_1, \tag{4.7}$$

$$z = a_1 x^2 + b_1 y^2 + c_1 xy + d_1 x + e_1 y + f_1, \tag{4.8}$$

A standard least-squares technique can be used to compute the coefficients.

Comparative experiment. We compare these methods with respect to their sensitivity to noise and sparseness. In order to test the sensitivity of the methods to noise, we add zero-mean Gaussian white noise to the available measurements of each of the synthetic images. The variance (σ^2) of the Gaussian noise used in our experiments varies from 0.01 to 0.1, and all experiments are performed for a density rate of 10%. For each technique we calculate the sensitivity to noise using the following formula:

$$difference\_of\_RMS = RMS(Z_n, GT) - RMS(Z_0, GT), \tag{4.9}$$

where, GT is the ground truth image, Z_n is the reconstructed surface from noisy measurements, n indicates the variance of the noise and Z_0 is the reconstructed surface from non-noisy measurements. The performance of each method for different amounts of noise is plotted in Figure 4.3.

In order to test the sensitivity of the methods to sparsity, for each of our synthetic images, we use different data constraint densities, varies from 5% to 20%. The average RMS reconstruction error for each method versus the sparsity is shown in Figure 4.4.

As can be seen in Figure 4.4, the thin-plate smoothing spline performed better than the other fitting methods in terms of sparsity. Therefore, in our application we chose to use the thin-plate spline as an interpolant for segmented regions. Figure 4.3, shows that the polynomial least squares fitting methods are adequate for normal noise, however, the closeness of their performance in terms of noise to the performance of thin-plate spline is acceptable. In addition to the results of our experiments, another motivation for using splines is that they offer a unified representation of surface information obtained from various visual cues such as stereo, that has been well argued by both Grimson [34] and Terzopoulos [83]. The other motivation is that the polynomial least squares fitting does not guarantee smoothness. The result of surface reconstruction on two sample synthetic images by the use of thin-plate spline as an interpolant has been shown in Figure 4.5, 4.6;

In order to validate the efficiency of our approach for discontinuity detection and preservation, we made a comparison of our result with the result of one of the common surface reconstruction methods, Terzopoulos [83], which dynamically adjust the discontinuity model, during surface reconstruction such that its continuity becomes consistent with discontinuities implied by data. Experimentally we found that the discontinuity detection and preservation performance of our approach is close to the Terzopoulos' (see Figure 4.7 for an example), however our technique performs much faster (the average computational time of our method for interpolating the synthetic images was three times less than the computational time of Terzopoulos' in the exact same conditions).

4.3.3 Refining the sparse disparity map

The results of our experiments on a set of synthetic data, presented in the previous section, provoked us to use thin-plate spline for interpolating image regions segmented by watershed transformation. However there are two main differences between the sparse disparity maps obtained from our stereo matching system, and the synthetic images used in our surface reconstruction experiments: (i) the obtained disparity maps are very sparse, often with density rate of less than 5%, (ii) the seed points in the obtained disparity maps are not distributed as evenly as they are in our synthetic test data. Therefore, in order to increase the number of seed points and make the distribution as even as possible, we employ a refinement technique based on the triangular and epipolar (described in section 3.4.1) geometrical constraints to yield a denser disparity map.

Our refinement algorithm begins with the sparse disparity map produced by our stereo matching system. Similar to [33], we add new corresponding points to the current set of matched points, using epipolar geometric and triangular constraints. Imposing these constraints helps us to prune the search space for feature point matching. The algorithm proceeds in the following steps:

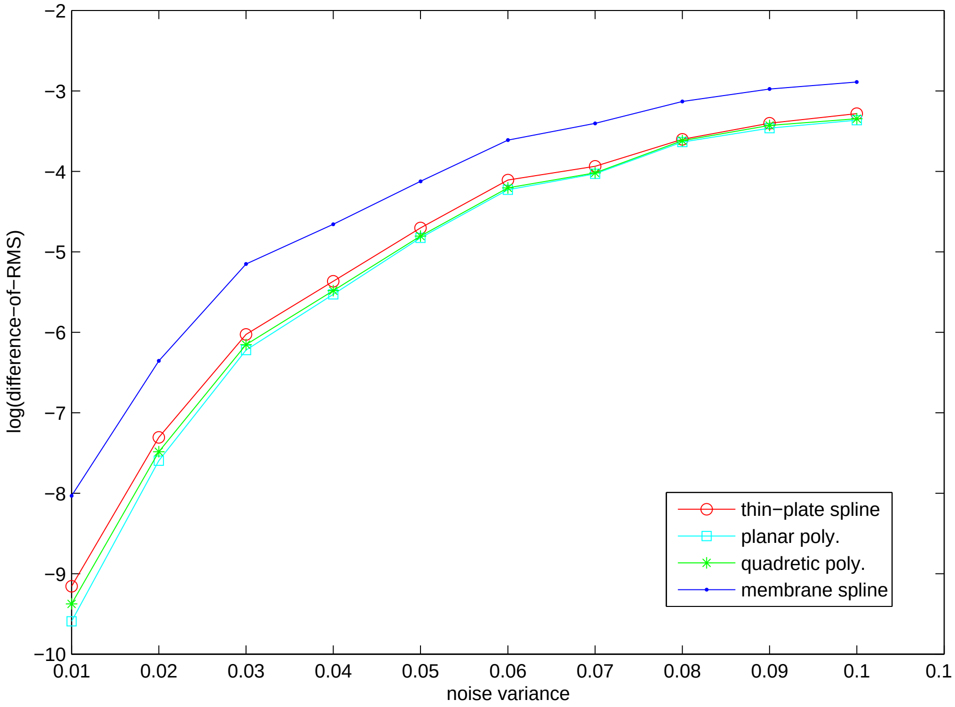


Figure 4.3: noise sensitivity of different interpolation techniques (thin-plate spline, membrane spline, planar and quadratic polynomials).

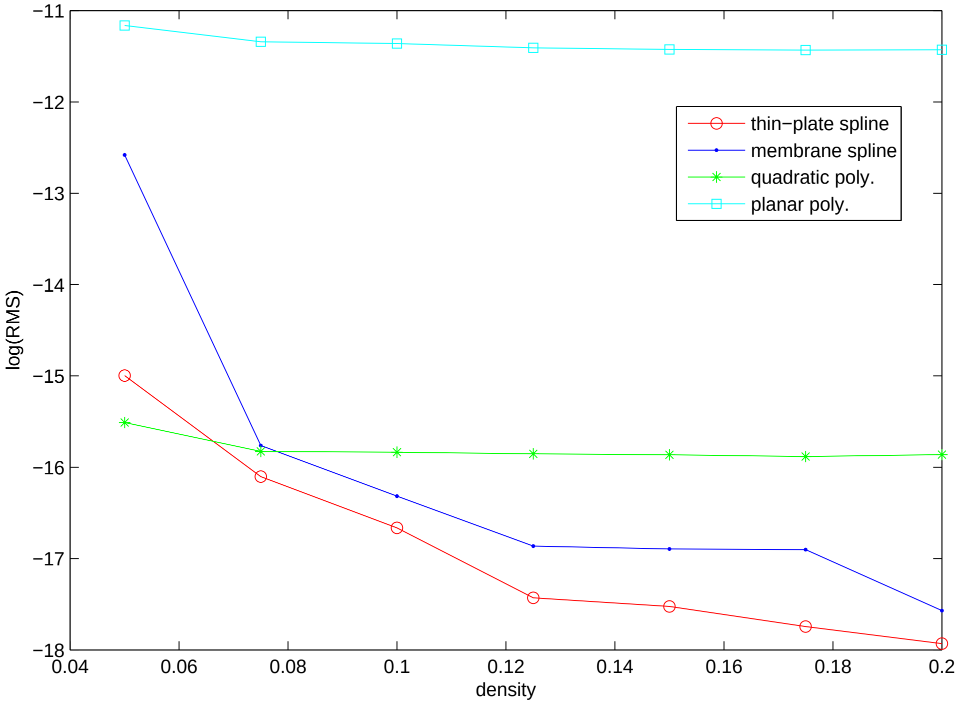


Figure 4.4: Average RMS error in surface estimation using thin-plate spline, membrane spline, planar and quadratic polynomials, as a function of measurements density (5%, 10%, 15%, and 20%).

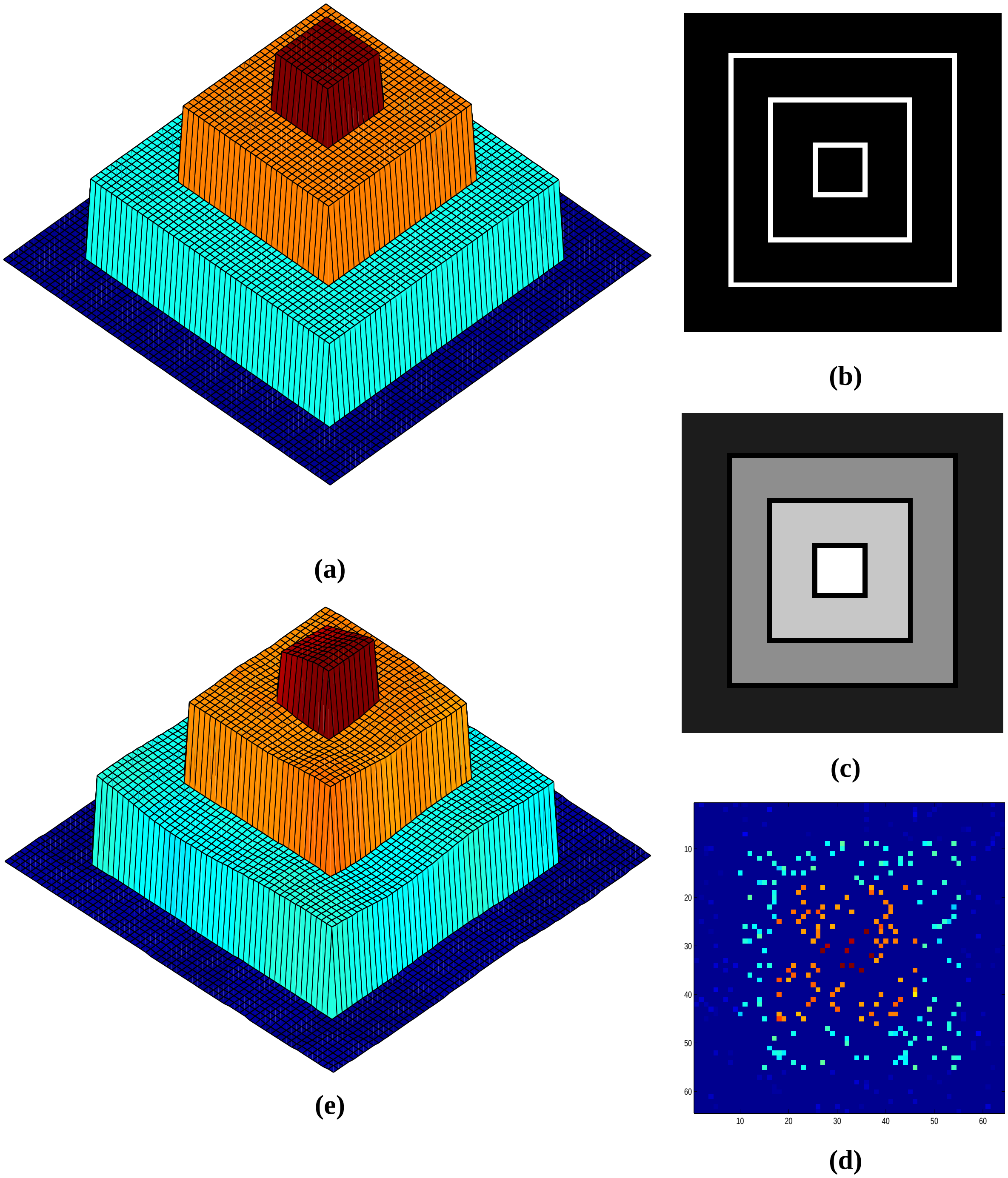


Figure 4.5: Reconstruction of a synthetic image from depth constraints. (a) an artificially generated ground truth (this is a standard artificial image used in many works, e.g., [83] and [34], for evaluation), (b) edge map obtained by phase congruency, (c) watershed segmentation of the image using the edge map, (d) constraints randomly sampled with density of 10% from (a) and normal noise with variance: 0.01, (e) reconstructed surface by thin-plate smoothing spline.

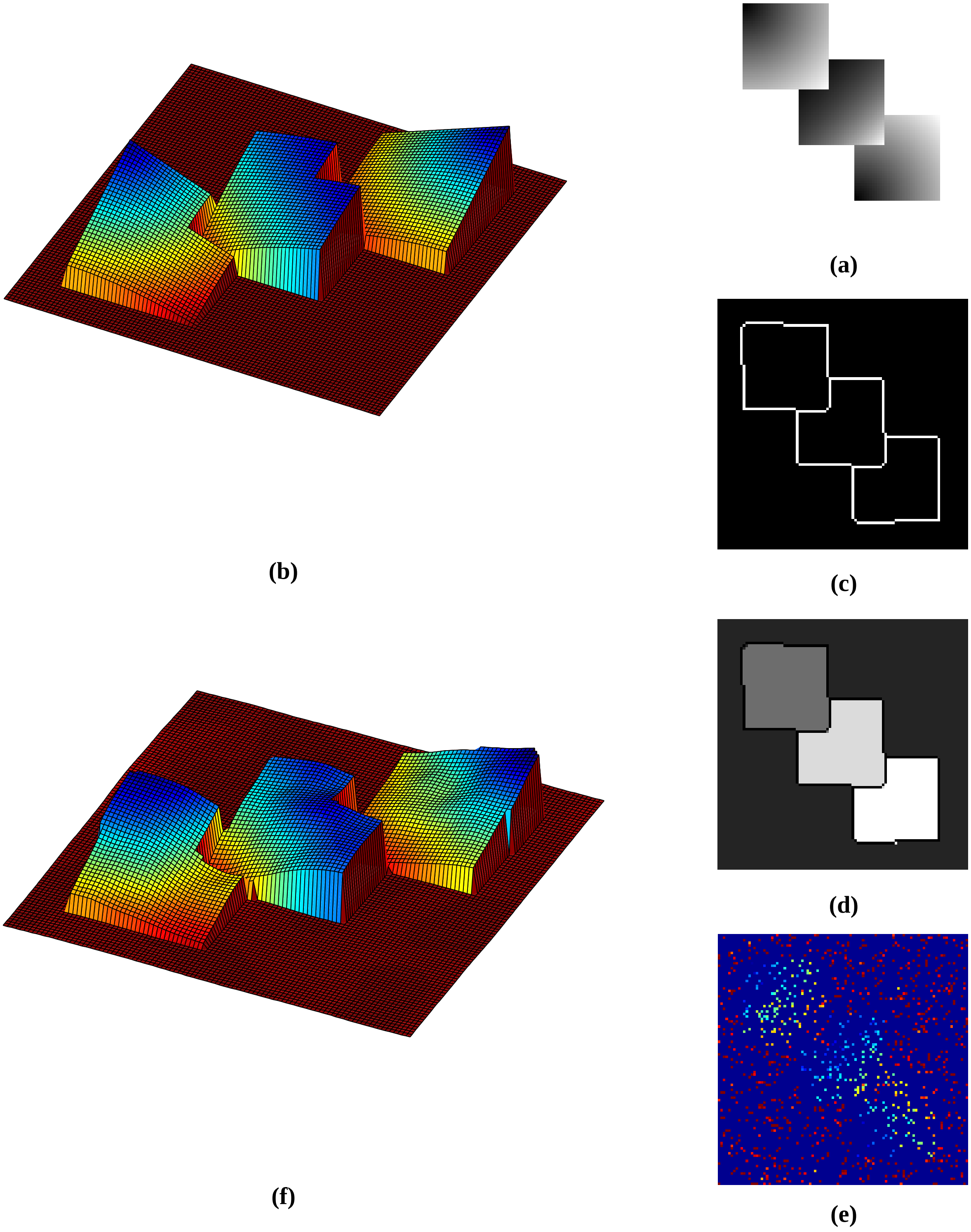


Figure 4.6: Reconstruction of a synthetic image from depth constraints. (a) an artificially generated ground truth (128x128 pixel), (b) 3D of (a), (c) binary edge map obtained by phase congruency, (d) watershed segmentation of the image using the edge map, e) constraints randomly sampled with density of 10% from (a) and normal noise with variance: 0.01, (f) reconstructed surface by thin-plate smoothing spline.

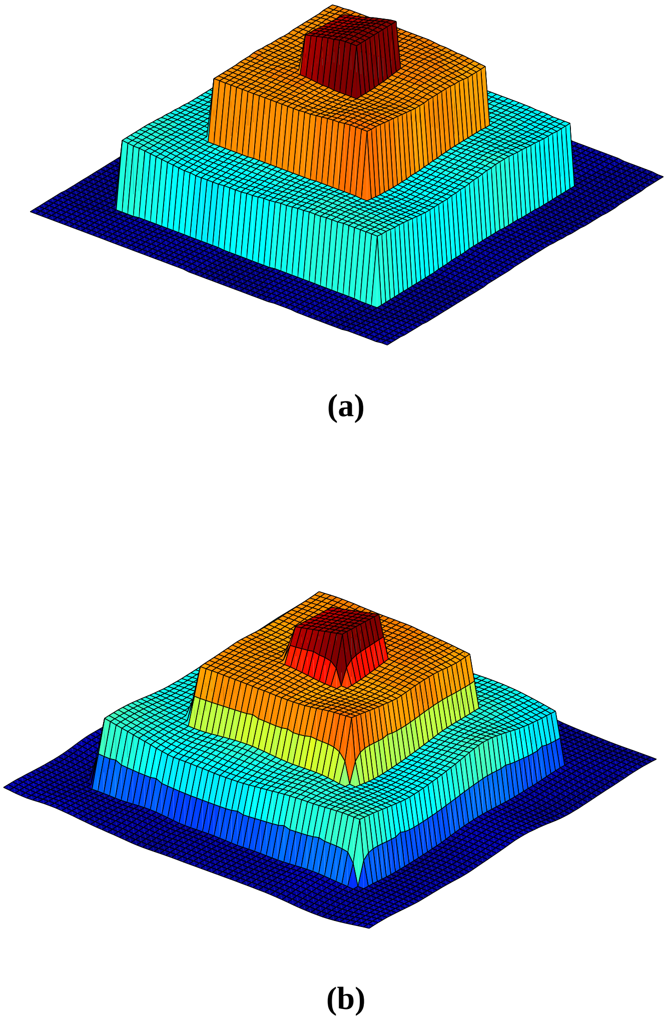


Figure 4.7: Reconstruction of the sparse disparity map in Figure 4.5.d.; (a) our method with log(difference-of-RMS) error = -9.18, (b) Terzopoulos' method with log(difference-of-RMS) error = -8.92; (density: 10%, noise variance: 0.01).

1. The Delaunay triangulation[5] on the matched points of the first image is computed and the triangulation of the second image is estimated from that: three vertices of a triangle in the second image are corresponding to the vertices of the same triangle in the first image.
2. For each triangle in the first image the center of the triangle, m_i, is considered as a potential feature point, if it is located at a textured area[6]. In order to find the possible matching partners in the second image, m'_j, we restrict the search area to those points which are located within the corresponding triangle in the second image and in a narrow band of width ε $(\varepsilon \in (2 \sim 4))$ pixels centered on the corresponding epipolar line, $l_i = Fm_i$; (see Figure 4.8). The distance between each candidate match, m'_j, and the epipolar line l_i is calculated using the Euclidean distance, based on the following equation:

$$d(m'_j, Fm_i) = \frac{|\, m'^T_j Fm_i \,|}{\sqrt{(Fm_i)_1^2 + (Fm_i)_2^2}}, \tag{4.10}$$

   where $(Fm_i)_k^2$ is the k^{th} component of vector Fm_i.
3. For each feature point m_i in the first image, the candidate feature, m'_j, in the second image which maximizes the similarity measure (Equation 3.26) is selected as the corresponding point.
4. Once the best match for each feature point in the first image is found, we use a threshold value to determine whether the similarity value between the descriptors of the points is high enough to consider them as a valid match.

This process stops when no more matched pair is added. Although by applying this algorithm on a sparse disparity map, the sparsity would be improved and measurements would be more evenly distributed, there are still some segments of the reference image that do not contain enough measurements. These segments either do not contain enough texture or suffer from partial occlusion, and therefore we do not reconstruct the surface of such segments.

[5] See Appendix A, for a brief introduction to Delaunay triangulation.

[6] To measure the texture at each location the entropy of the normalized histogram of pixel values within a window around that location, was used. This entropy value should be greater than a predefined threshold value.

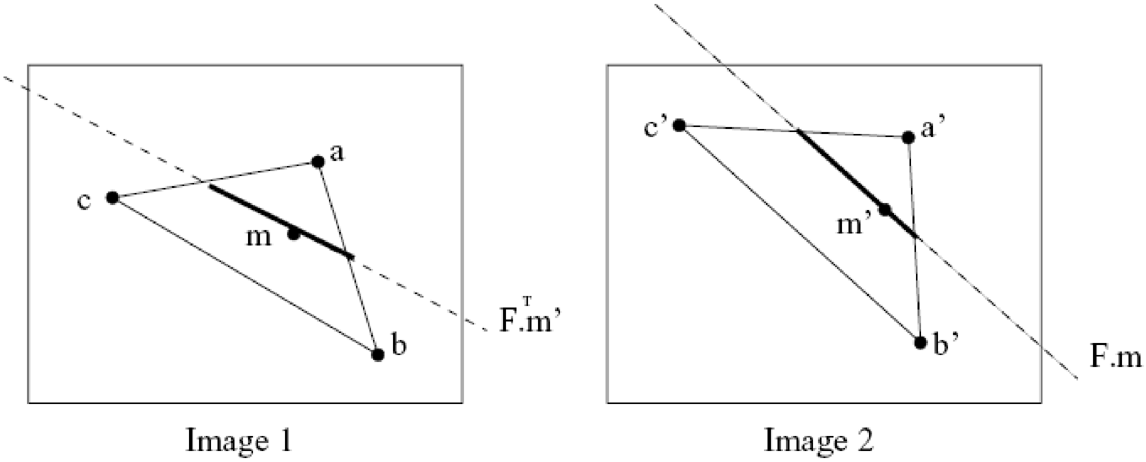


Figure 4.8: (Reproduced from [33]). The correspondence of the point m in the first image within the triangle (a,b,c), is the point m' in the second image, located on the epipolar line $F.m$ within the triangle (a',b',c'). Similarly, the point m' has its correspondence in Image1 on $F^T.m'$ and within (a,b,c);

4.4 Chapter summary

In this chapter, we described our method to compute dense disparity map from infrared stereo image pairs. First, the disparity maps produced by our stereo matching method (described in the previous chapter) are refined in terms of density rate and measurements distribution, using triangular and epipolar geometrical constraints. In order to densify sparse disparities we developed a surface reconstruction technique which makes use of the prior knowledge of depth discontinuities in the reconstruction process. Reconstruction is performed separately in image regions segmented by watershed method. Our analysis of some of the surface fitting methods on synthetic data prompted us to use thin-plate splines which are more robust in the face of high sparsity and noise.

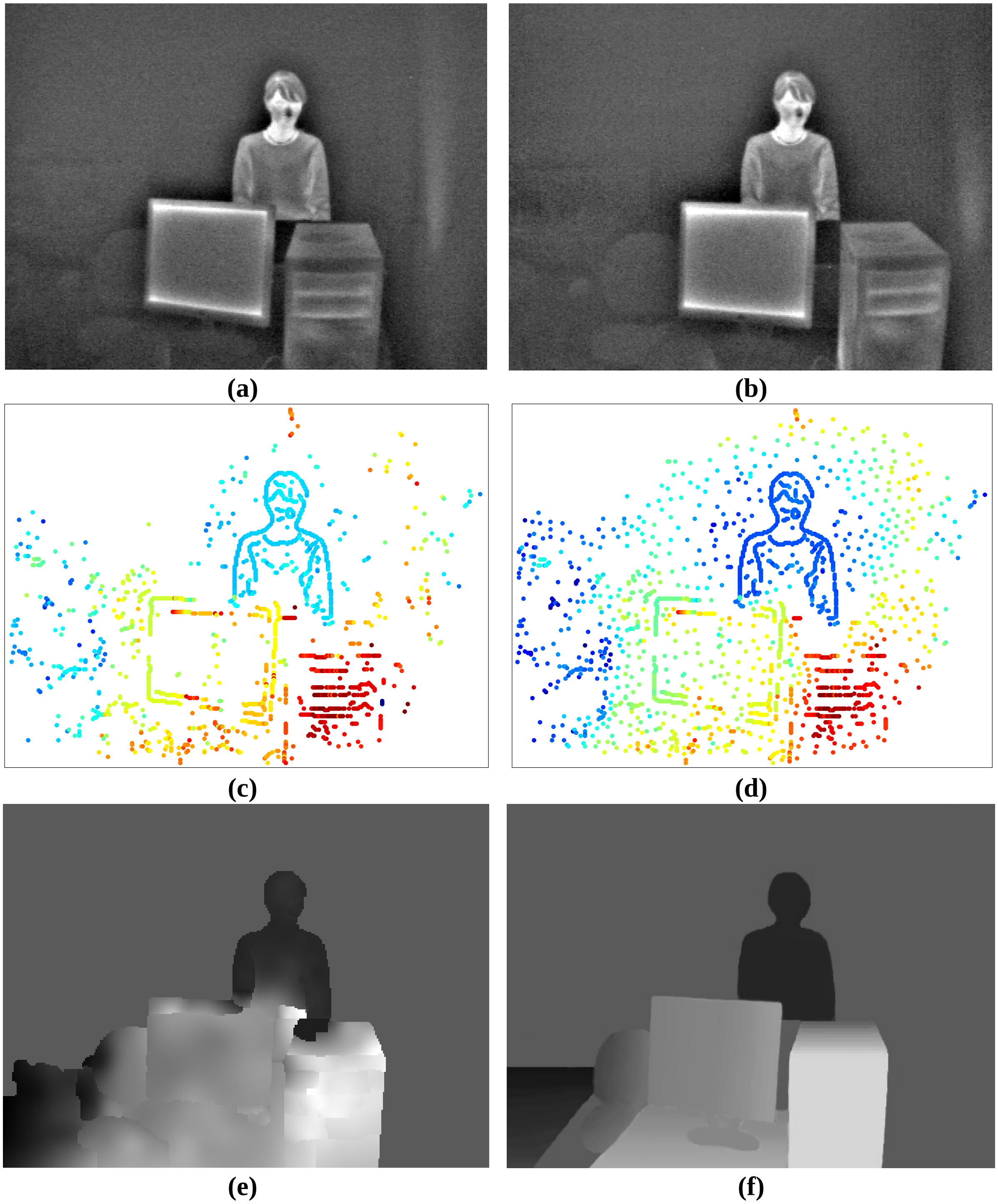


Figure 4.9: (a) left image of an IR stereo pair (reference image); (b) binary edge map; (c) sparse disparity map; (d) refined sparse disparity map; (e) produced dense disparity map; (f) ground truth.

Chapter 5

Experimental results

Having explained the structure and operation of our technique, we now evaluate its performance, with regards to a variety of measurements, on a dataset of several infrared stereo images. Our experiments consist of two main stages: in the first stage, the intention is to test the performance of our feature matching technique on a set of indoor IR pairs and compare it quantitatively with the performance of several state-of-the-art feature matching methods widely used for visible images; in the second one, the goal is to validate the efficiency of the our technique in producing dense disparity maps. Although the main aim of this thesis is to extract structure from indoor IR images, we are also interested in investigating how well our matching model performs on outdoor images; therefore, a qualitative assessment of our technique is performed for outdoor IR images, as well. In the following, we first describe the dataset used in our experiments and then we present and discuss our experimental results.

5.1 Experimental setup

There is not any publicly available dataset of infrared stereo images that we can use for testing our method. However, there is one database of infrared images of which two sets [2] (one consisting of IR & color pairs, and the other is a motion sequence of a single IR camera) are used to evaluate our results in addition to our own IR stereo image data. Our IR image pairs were created by setting up a stereo configuration of two commercial IR cameras mounted on a commercial Manfrotto stereo tripod system (as can be seen in Figure 5.1.a and 5.1.b). Each IR camera used in our setup is a Raytheon ControlIR 2000b camera which is an uncooled ferroelectric type with the resolution of 320 by 240 and spectral response of 7-14 μm. The type of detector used in the cameras is a hybrid ferroelectric staring focal plane array and is utilized by pyroelectric and dielectric effect measuring. The sensor material is Barium Strontium Titanate (BST). The startup time required for the cameras is less than 30 seconds typical and less than 90 seconds maximum. Each camera is outfitted with an 18 mm lens and the field of view is 45 degrees by 35 degrees. The depth of field is 5 m to infinity (at infinity focus) and the focus range is 3 m to infinity. The adjustable iris permits f/1.0 to f/8.0. The input voltage to operate each camera is 9 to 28 VDC with operating current less than 1 A at steady state and less than 7 A at start-up (ambient for 20 μs). The weight of each camera with the lens is 1.6 lbs.

Using our IR stereo configuration, we captured several IR stereo pairs from indoor scenes. All acquired images are 320×240 pixels, and pre-processed using intensity adjustment (remapping intensity values to the specified range of [0 255]) and Gaussian low-pass filtering (of size [3 3] with $\sigma = 0.5$) in order to reduce the noise influence (some sample IR images from our dataset are shown in Figure 5.2). For each stereo pair, a ground-truth disparity map is also constructed for the foreground objects of the scenes to be used in our experiments. The reason that we only consider the disparity of foreground objects in our quantitative evaluation is that background objects in indoor environments are usually at "thermal crossover" (i.e., thermal properties of the objects are relatively similar to those of the surrounding environment [17]), and therefore the captured IR images at those areas do not contain enough information for image processing tools.

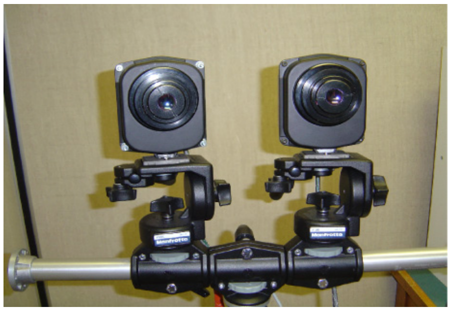


Figure 5.1: The stereo rig; a close up of the dual IR cameras mounted on a tripod.

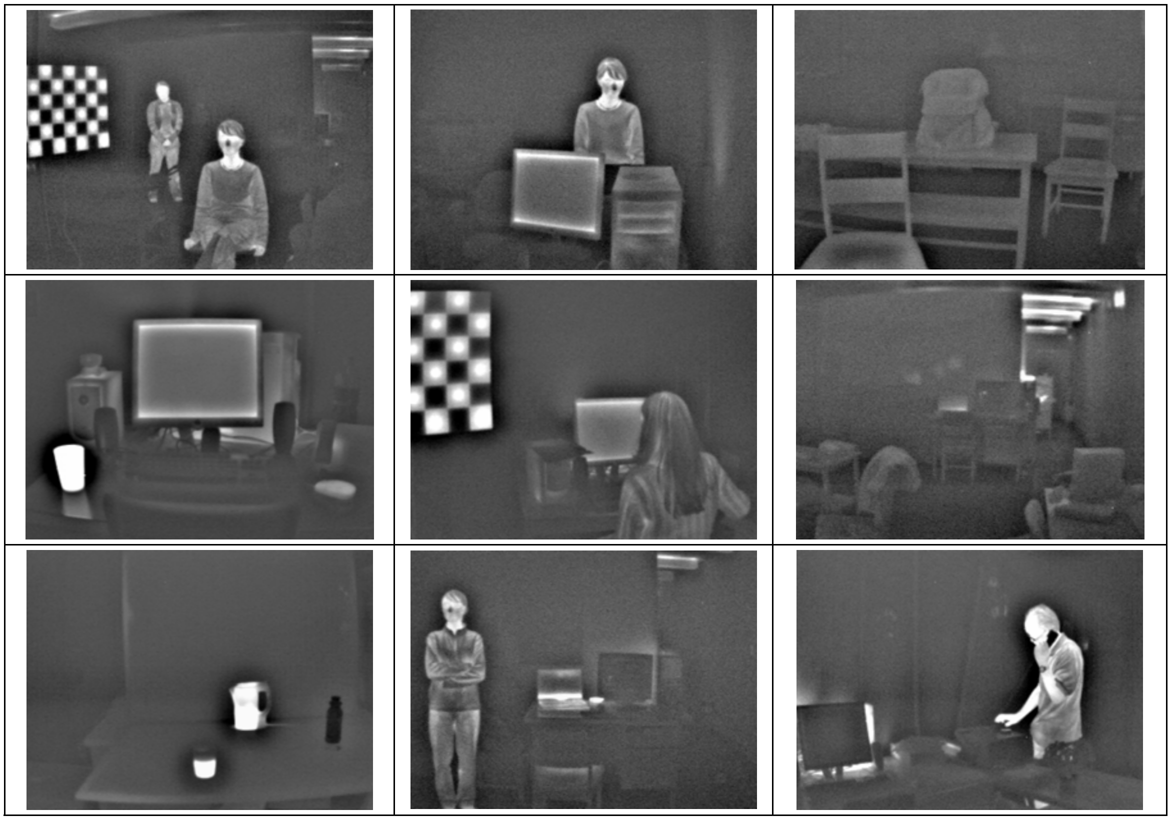


Figure 5.2: Some samples of the left images of our IR stereo pairs.

5.2 Results

5.2.1 Feature matching

In order to be able to validate the efficiency of the feature matching technique used in our method and compare it with other methods quantitatively, we perform a comparative experiment on a small subset of our IR stereo dataset. Besides our feature matching method, the other feature matching techniques used in our comparative experiments are Harris-NGC[1], DoG-SIFT, Canny-NGC, and KLT (all these techniques have been described in Chapter 3). We applied all techniques on our test stereo images, and for each method, calculated the average results in terms of the number of feature points detected in the left and right images, and the mismatch percentage. In our experiments, similar to other works in the field (e.g. [29]), we calculated mismatch percentages by manually counting the number of correct

[1] Normalized Grayscale Correlation

matches using a 5×5 window, for each test stereo pair, and dividing it by the total number of matches. Table 5.1 summarizes the results.

As can be seen, although most of the feature matching techniques used in our comparative studies are among the widely used matching methods for visible images, their results on our IR stereo test images are not well enough, and our matching method achieved the best average results both in terms of the number of detected features and the mismatch rate. In our method, in average, 3685 features were detected in the left images and 3425 features in the right images. 78% of detected features were matched and the other features were discarded due to occlusion or left-right consistency constraint; among the matched points, 14% were outliers. Using our matching refinement technique based on RANSAC fitting, the final mismatch rate reduced to 1.5%.

Method			**Result**			**Used in**
Detector	**Descriptor**	**Matching Metric**	**Nbr of Feature Pts**	**Matched points**	**Mismatch %**	
Harris	Normalized Grayscale Correlation		307	240	21 %	[71]
DoG	SIFT	Euclidean distance	211	65	10.3 %	[57]
Canny	Normalized Grayscale Correlation		4481	2241	17 %	[80]
Kanade-Lucas-Tomasi (KLT)			-	99	15.5 %	[74]
Our technique without matching refinement			3555	2772	14 %	
Our technique with matching refinement			3555	2178	1.5 %	

Table 5.1: Compared efficiencies of the matching methods described in Chapter 3.

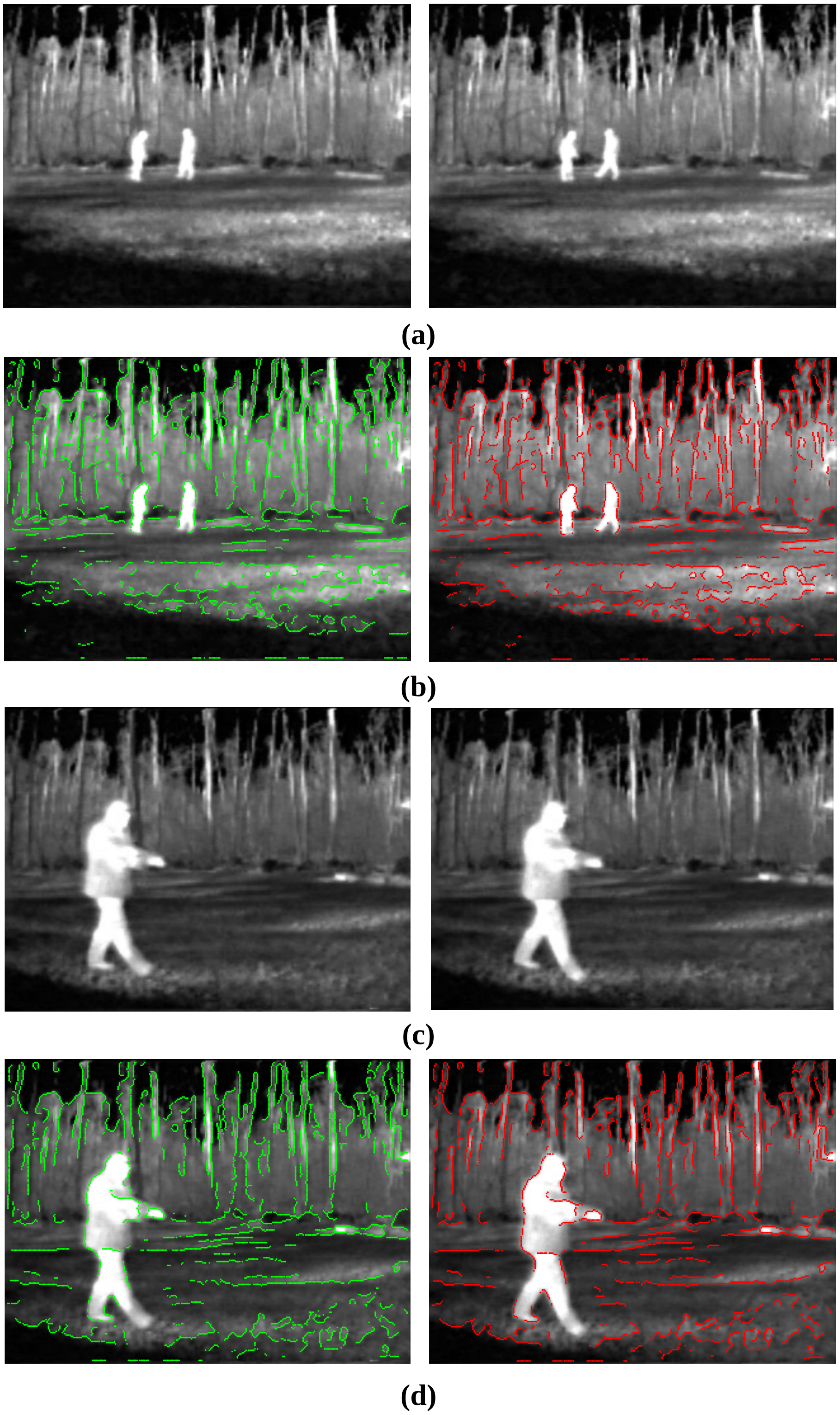
5.2.2 More results on feature matching

In the previous section we have analyzed the performance of our matching algorithm on our dataset of IR stereo images. However, since all the images in our dataset are from indoor scenes, we are interested in investigating how well our matching model performs on outdoor IR images. Furthermore, we would like to discover how similar are the features matched from IR images of a scene, to the features matched from the visible images of the same scene (to see if fusion of visible and IR imaging can perhaps be a potential direction for improving the results). To address these investigations, we use a publicly available database, OTCBVS [2], that is a collection of IR infrared and visible images and videos. We use two of the seven datasets available in OTCBVS in our evaluations, namely: the Terravic Motion IR Database and the OSU Color-Thermal Database. Although these databases do not contain stereo images, all the data sequences are of motion and can be used for matching purposes as test.

To investigate the performance of our matching on outdoor IR images, we have applied the method on several pairs of frames from the outdoor sequences of the Terravic Motion IR Database. Analyzing the results qualitatively, we found out that the performance of our matching on outdoor images is almost as good as it is on indoor images (see Figure 5.3 for some examples). Furthermore, we figured that the average number of matched points detected in outdoor test pairs was higher than the average number of matched points detected in indoor test pairs (as can be observed by comparing Figure 5.3.b and 5.8.b., for example). This could be due to the higher complexity of the outdoor scenes which is because of the fact that outdoor environments are more dynamic (both visually and thermally) and therefore the number of the objects (scene areas) which are at thermal crossover is less in outdoor scenes than it is in indoor scenes.

To exploit the similarity between the matched features from IR images of a scene, and the matched features from the visible images of the same scene, we have applied our matching method on several IR pairs and their corresponding visible ones, from the OSU Color-Thermal Database. Figure 5.4 shows the result of applying our matching algorithm on these pairs. As can be seen there is a large overlap between the matched points detected in IR pairs and those detected in visible pairs. But currently, with our algorithm, matching the features from the IR and color images is not possible as the feature descriptors in visible and IR domains are not close.

The sudden illumination changes and the presence of shadows (see Figure 5.5 and 5.6), which can result miss-matches in visible pairs, are not issues in matching IR pairs, and on the other hand, the halo effect which is a problem in matching IR pairs, is not an issue in matching visible pairs. Hence, one can conclude that the fusion of visible and IR results can be a potential direction for improving the final performance of the system.



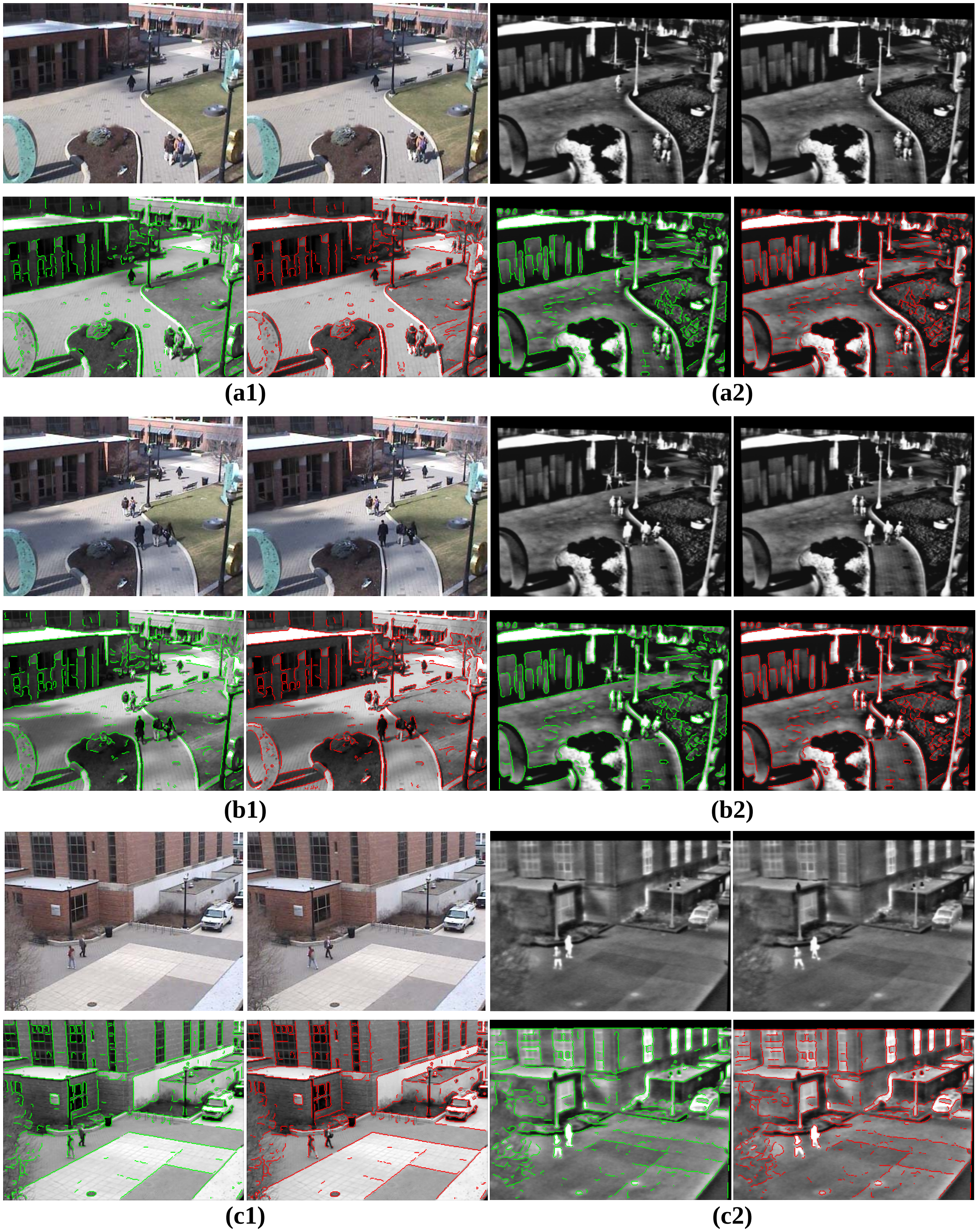
(a)

(b)

(c)

(d)

Figure 5.3: Some examples of feature matching for two pairs of outdoor IR images; (a), (c): frames taken from an IR sequence in dataset 05/otcbvs collection; (b), (d): matched features of (a), (c), which are illustrated in green & red in the left and the right images, respectively.



(a1) (a2)

(b1) (b2)

(c1) (c2)

Figure 5.4: Some examples of feature matching on three visible pairs (a1), (b1) and (c1), and their corresponding IR pairs (a2), (b2) & (c2). Frames have been taken from dataset 3/otcbvs.

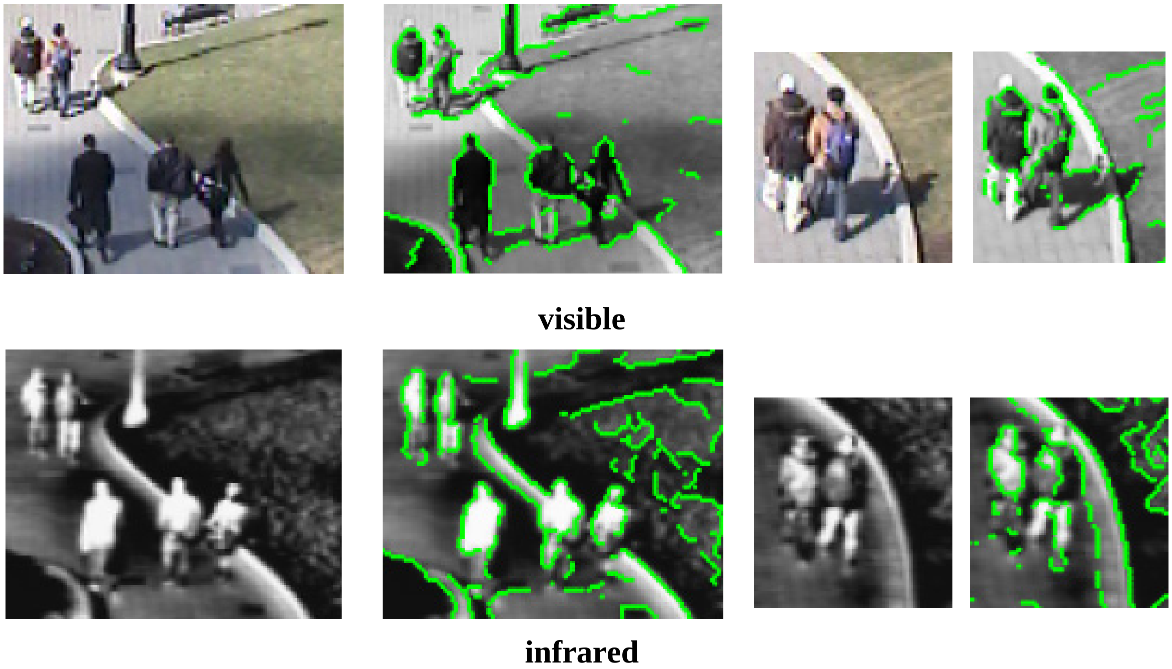


Figure 5.5: Some examples of sudden illumination changes/presence of shadows (pedestrians and the lamppost's shadows).

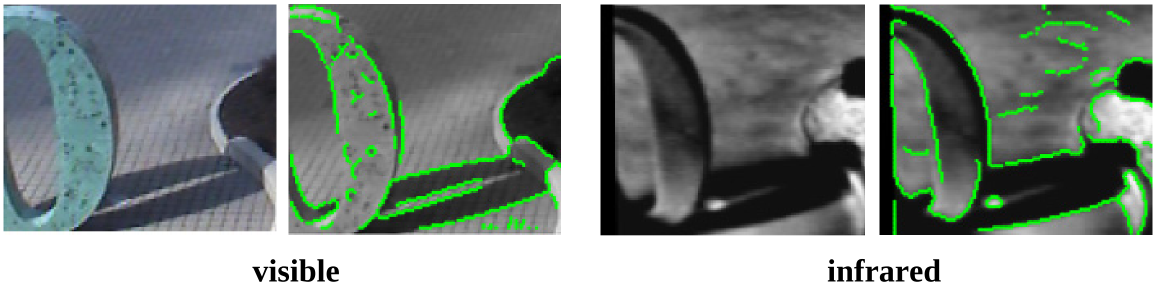


Figure 5.6: Some examples of the presence of shadows (statue shadow) in visible and infrared images; as can be seen from the IR images, the shadow has been stationary long enough to cool the background.

5.2.3 Reconstructed disparity map

We evaluated the disparity maps produced by our method, using our dataset of IR stereo images described in Section 5.1. For each stereo pair, the disparity map produced by our method was compared to the ground truth (similar to [81], [99]) and the error was computed by counting the number of pixels with disparities that differ by more than 5 from the ground truth, called bad matching pixels. In our statistics, we only take into account the disparities computed for foreground objects (as discussed in Section 5.1) and we also ignore occluded areas. The percentage of bad matching pixels is computed through the equation (5.1):

$$B = \frac{1}{N} \sum_{(x,y)} (| d_C(x,y) - d_T(x,y) | > \delta_d), \tag{5.1}$$

where N is the total number of pixels, $d_C(x,y)$ is the disparity map, $d_T(x,y)$ is the ground truth map and δ_d is the disparity error tolerance. For our experiments we set δ_d = 5 (similar to [81]).

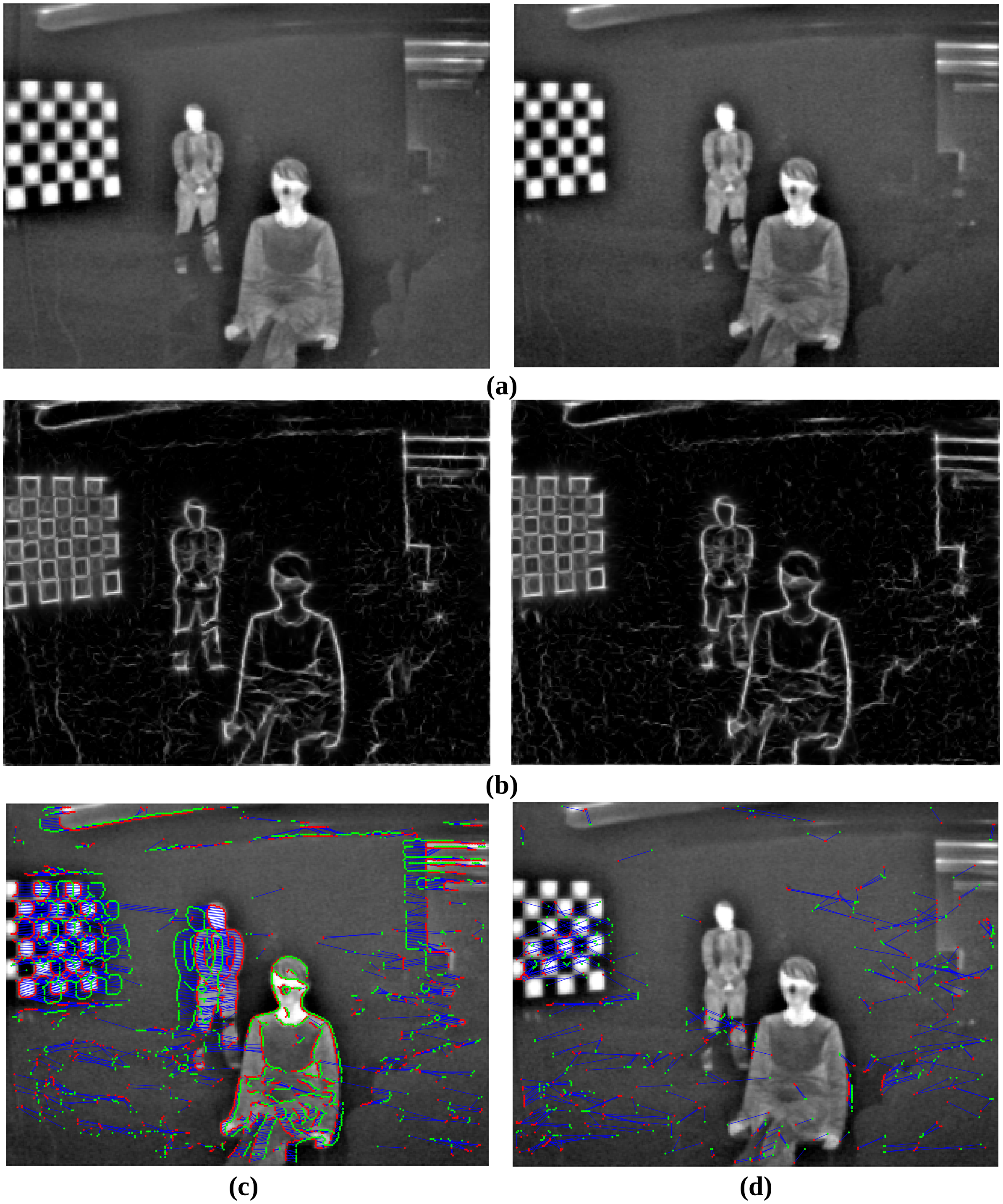
Our method achieved an average error percentage of 11%, which is comparable to the results achieved by state-of-the-art techniques in visible domain [1]. Results from different steps of our technique on four sample IR stereo pairs are illustrated in Figures 5.7-5.18. Figure 5.19 shows the result of the method on a sample of IR outdoor images.

Analyzing the results on the test set, we observed that a large percentage of the failure responses of our technique occurred in the regions where the temperature (intensity) is uniformly distributed. Furthermore, the presence of the halos in IR images (some samples of IR halos have been marked in green, in Figures 5.8.a, 5.14.a and 5.17.a) severely impairs the performance as the halo artifact around foreground objects occlude the thermal pattern (texture) of the objects nearby (halos can be detected and filtered out as a suggestion for future work). Also, the halo effect is minimized if objects are further away from the camera which they will be for outdoor applications

The computational time of the proposed method does not depend considerably on the content of the stereo images. The most computationally intensive components of the algorithm are the feature extraction and description, which take almost half of the computational time. The surface reconstruction can also be costly to compute, however as it is only applied to a relatively small number of segments (rather than all image segments), it does not significantly contribute to the overall processing time. Using un-optimized Matlab code on a 3.2 GHz Pentium 4 computer, we experienced typical processing times of 20-25 seconds for each stereo pair, depending on the complexity of the images. However, the computational time can be significantly reduced by using a faster programming language, such as C++, to implement the method. Also, by reducing the size of the images from 320×240 to 160×120 and 80×60, we can speed up the processing time by almost 66% and 83%, respectively, but at the expense of losing accuracy.

5.3 Chapter summary

In this chapter we evaluated the performance of our method, with regards to a variety of measurements, on a dataset of several infrared stereo images. Two main sets of experiments were performed while in the first one the aim was to validate the efficiency of the feature matching technique used in our method on a set of indoor IR stereo pairs and compare it with other methods quantitatively, and in the second one the goal was to evaluate the disparity maps produced by our method. In addition, some qualitative assessments of outdoor IR and color sequences have been done to investigate how well our matching model performs on outdoor IR images. The results are quite convincing and suggest that the concepts presented so far are feasible and are worthwhile to be used in real-world applications such as pedestrian detection and tracking for surveillance systems, robot obstacle detection in dark environment, passenger pose estimation for airbag systems, etc. A more detailed analysis of the possible improvements of the current work and the future directions will be provided in the next chapter.



(a)

(b)

(c)

(d)

Figure 5.7: (a) shows the original IR stereo pair; (b) the phase congruency edge maps; (c) displays the right image overlaid with the inliers of the right image (red points) and of the left image (green points), resulted from the matching refinement, along with the correspondences (with blue lines); (d) displays the detected outliers.

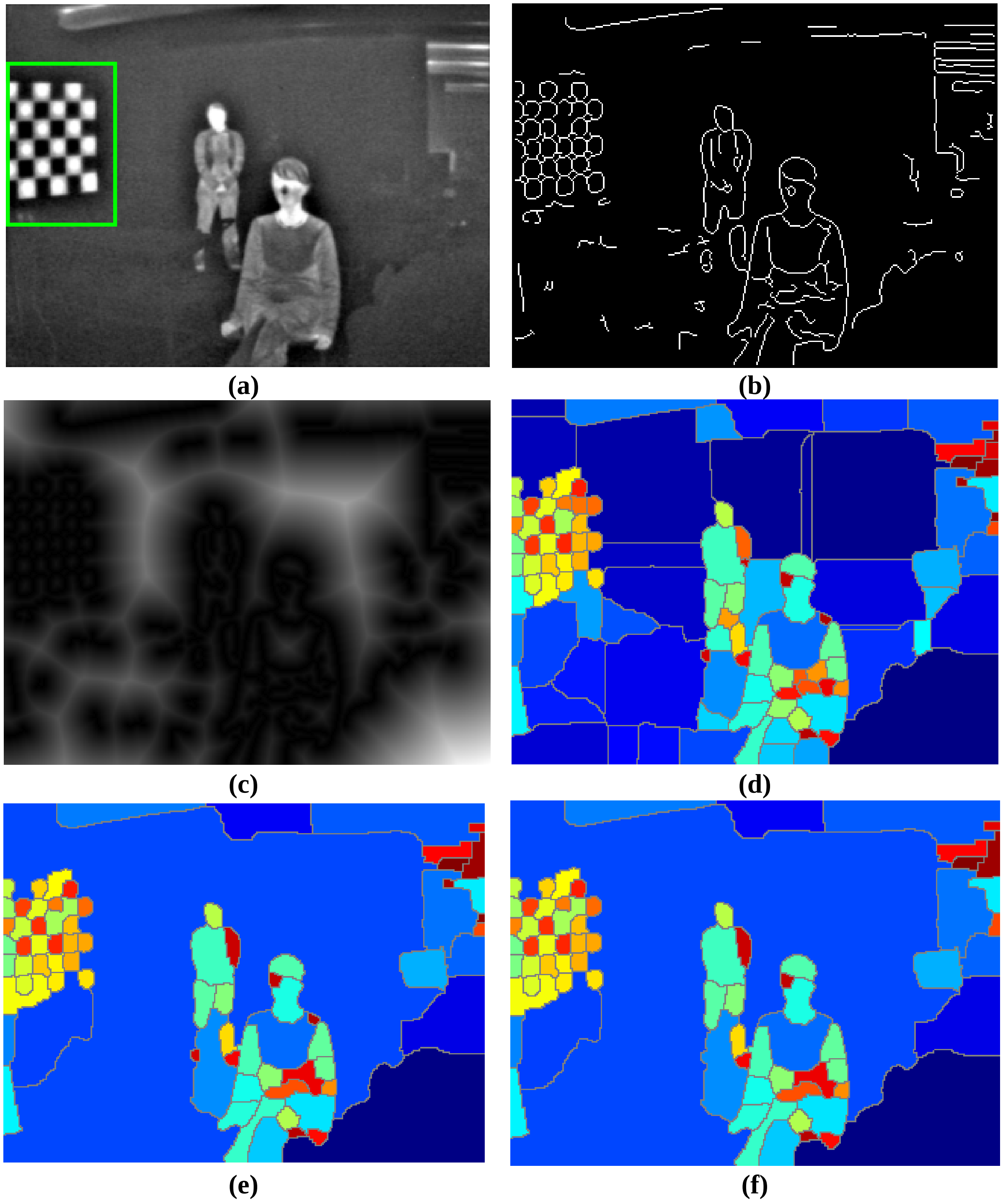
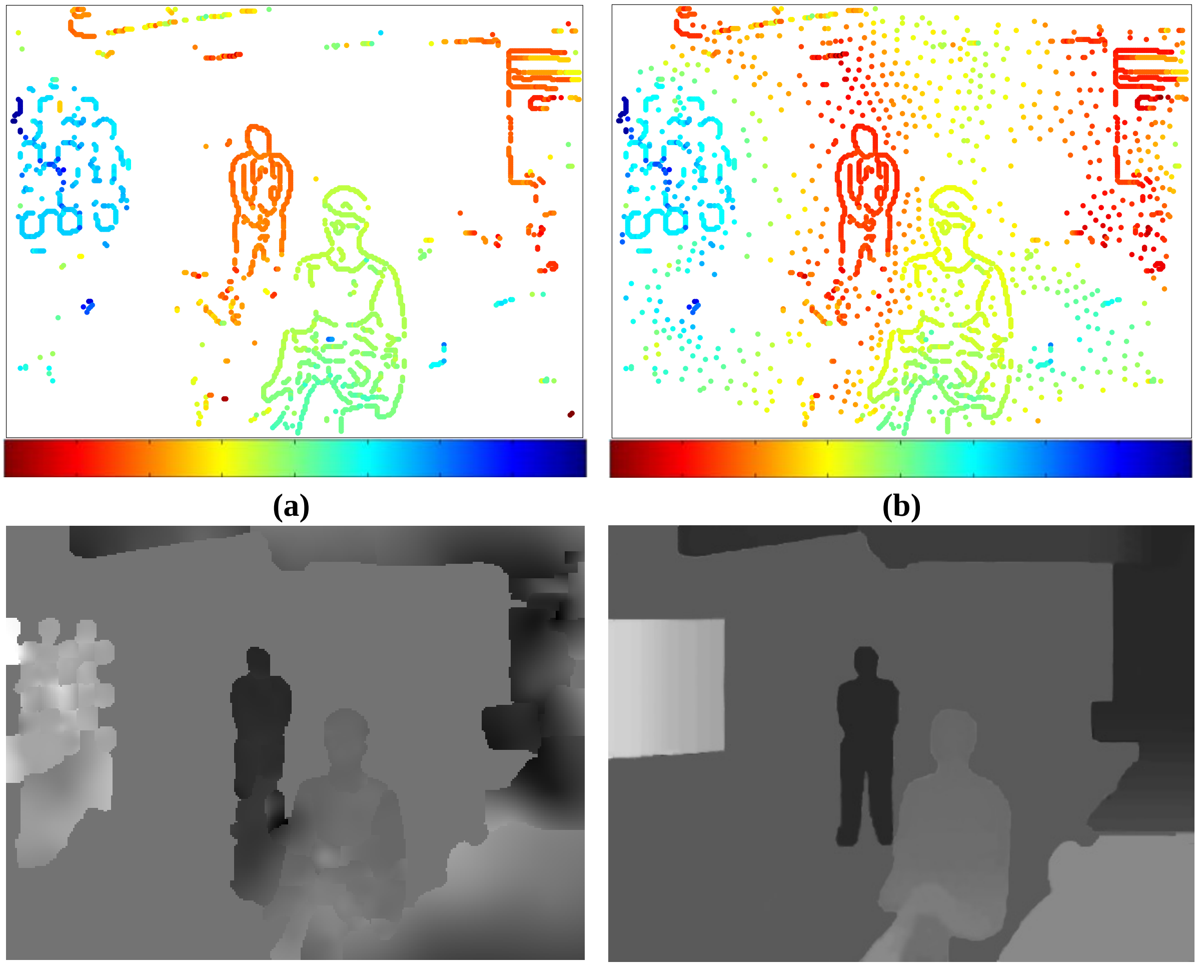

(a)
(b)
(c)
(d)
(e)
(f)

Figure 5.8: (a) right image of an IR stereo pair (reference image); (b) binary edge map; (c) distance transform of the binary map; (d) initial segmentation result; (e) segmentation result after removing irrelevant watershed lines; (f) segmentation result after removing irrelevant watershed lines and removing/merging small regions.



(a)

(b)

(c)

(d)

Figure 5.9: (a) sparse disparity map; (b) refined sparse disparity map after Delaunay triangulation; (c) produced dense disparity map; (d) ground truth.

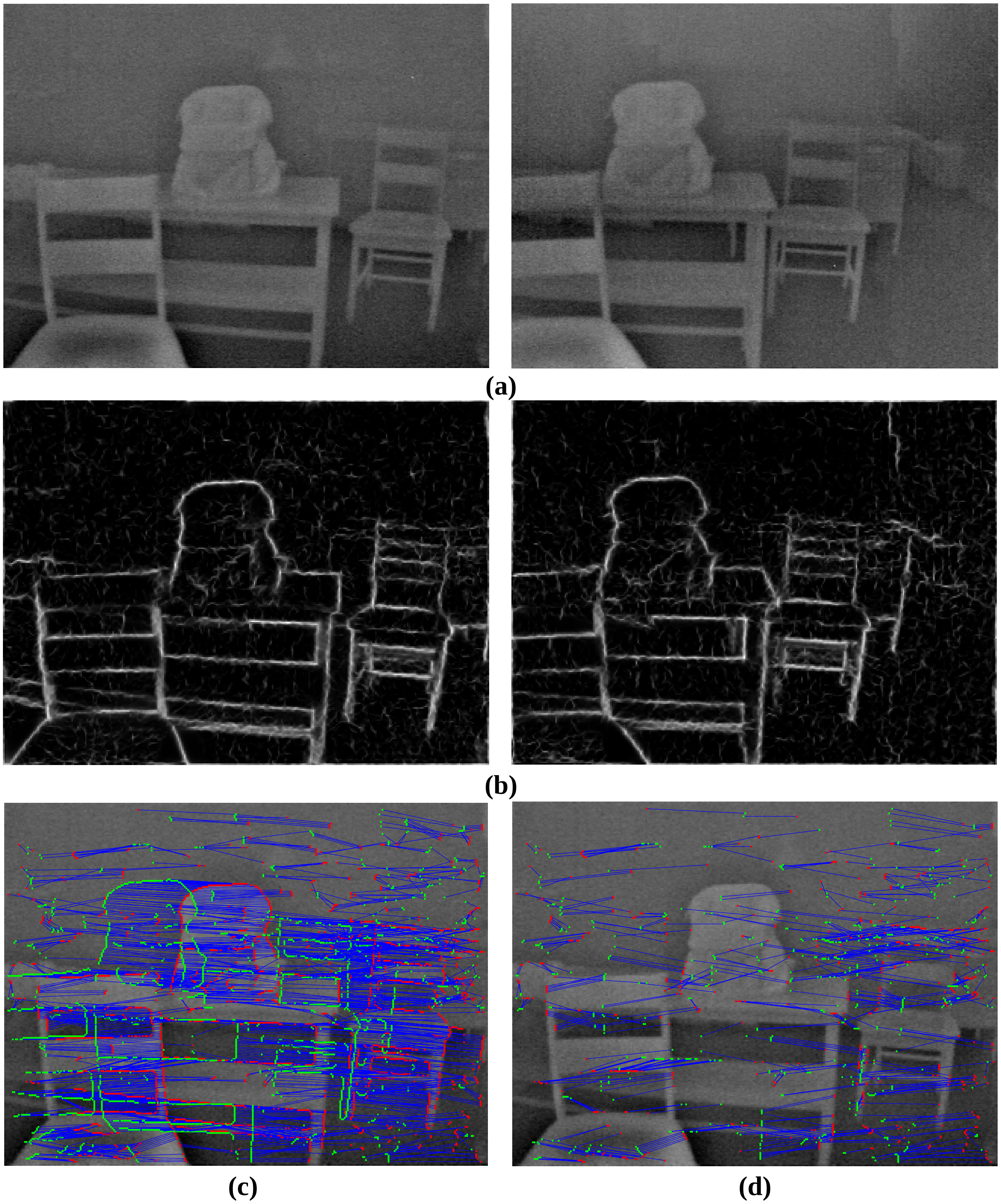


Figure 5.10: (a) shows the original IR stereo pair; (b) the phase congruency edge maps; (c) displays the left image overlaid with the inliers of the left image (red points) and of the right image (green points), resulted from the matching refinement, along with the correspondences (with blue lines); (d) displays the detected outliers.

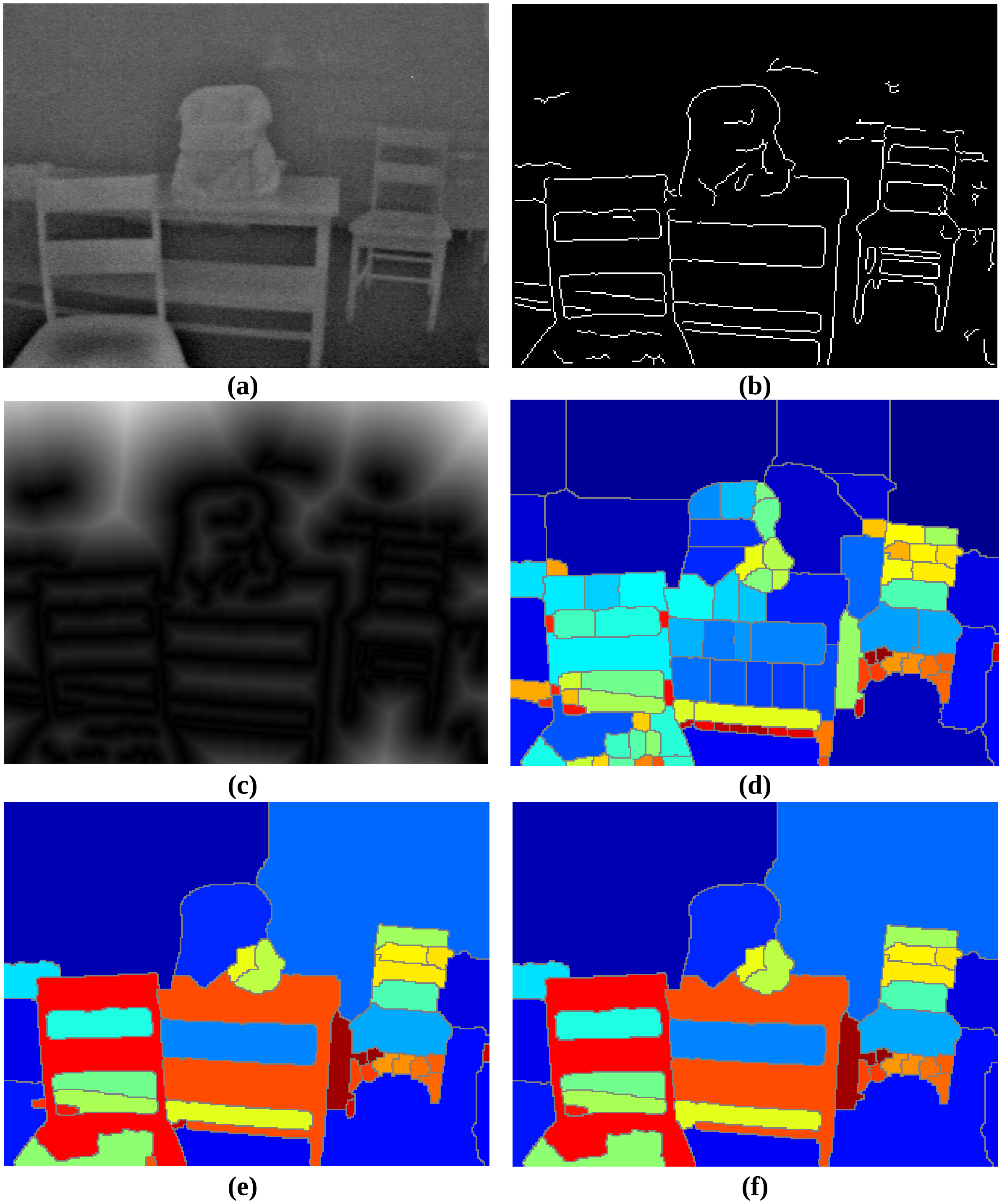


Figure 5.11: (a) left image of an IR stereo pair (reference image); (b) binary edge map; (c) distance transform of the binary map; (d) initial segmentation result; (e) segmentation result after removing irrelevant watershed lines; (f) segmentation result after removing irrelevant watershed lines and removing/merging small regions.

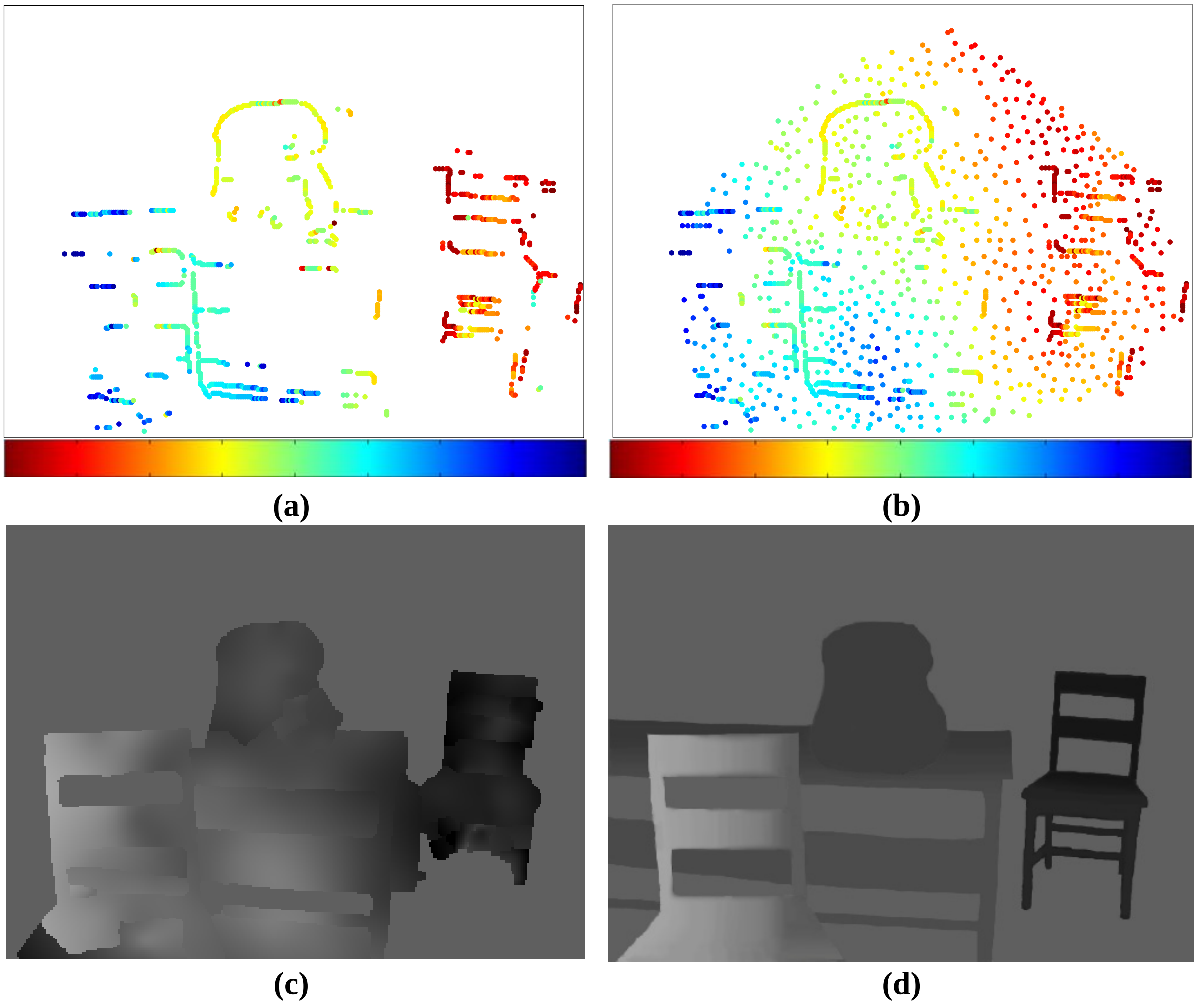


Figure 5.12: (a) sparse disparity map; (b) refined sparse disparity map after Delaunay triangulation; (c) produced dense disparity map; (d) ground truth.

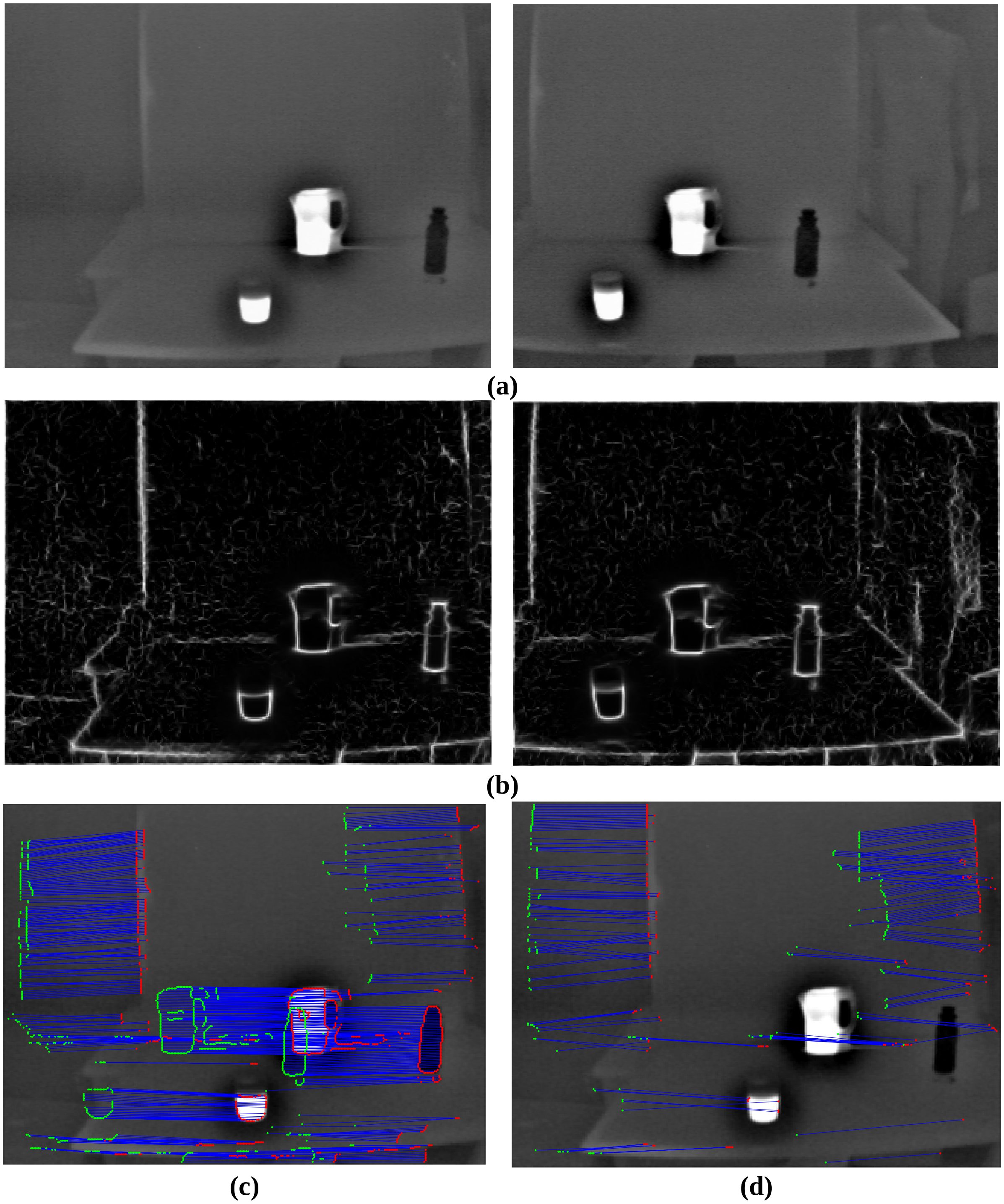


Figure 5.13: (a) shows the original IR stereo pair; (b) the phase congruency edge maps; (c) displays the left image overlaid with the inliers of the left image (red points) and of the right image (green points), resulted from the matching refinement, along with the correspondences (with blue lines); (d) displays the detected outliers.

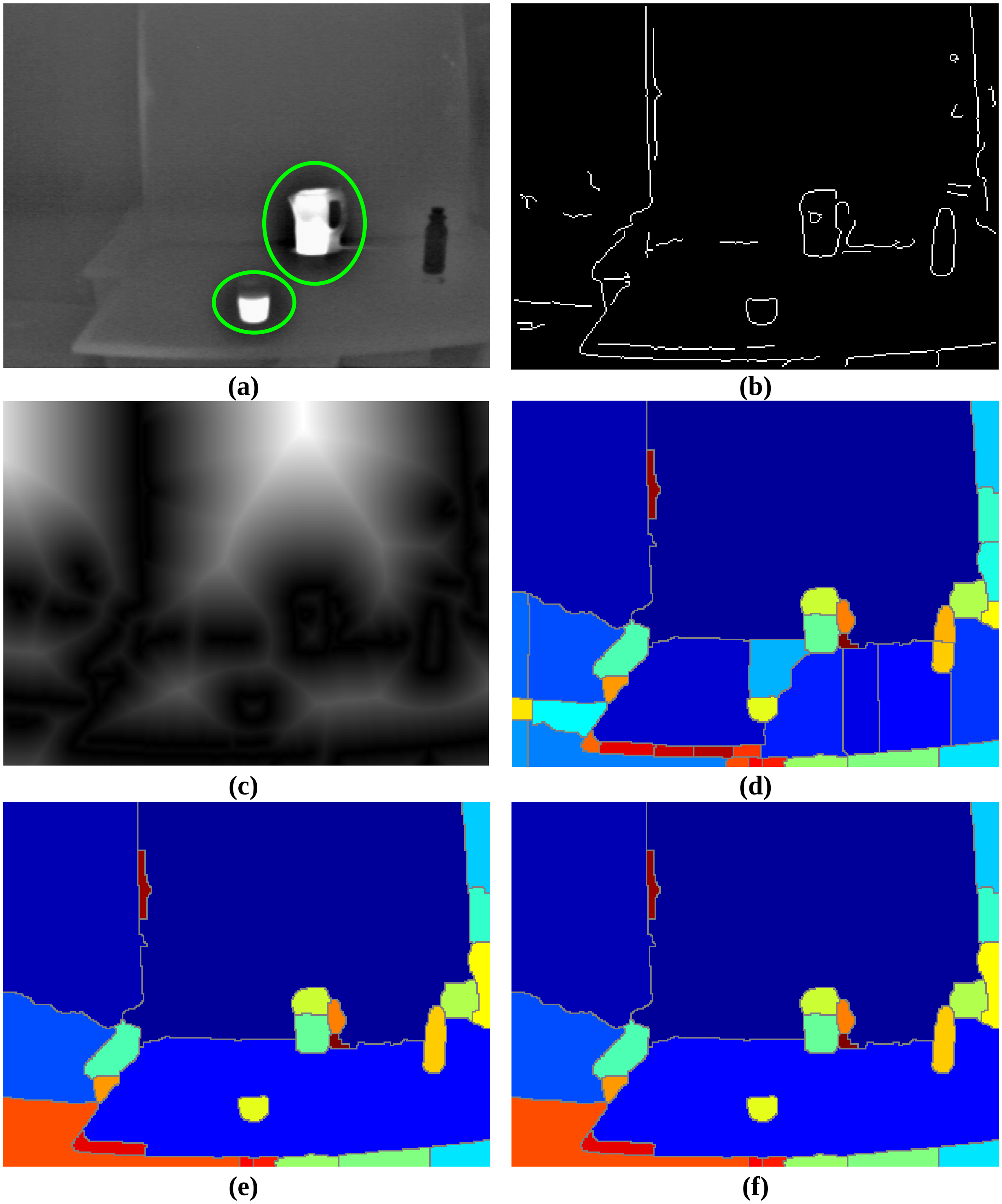


Figure 5.14: (a) left image of an IR stereo pair (reference image); (b) binary edge map; (c) distance transform of the binary map; (d) initial segmentation result; (e) segmentation result after removing irrelevant watershed lines; (f) segmentation result after removing irrelevant watershed lines and removing/merging small regions.

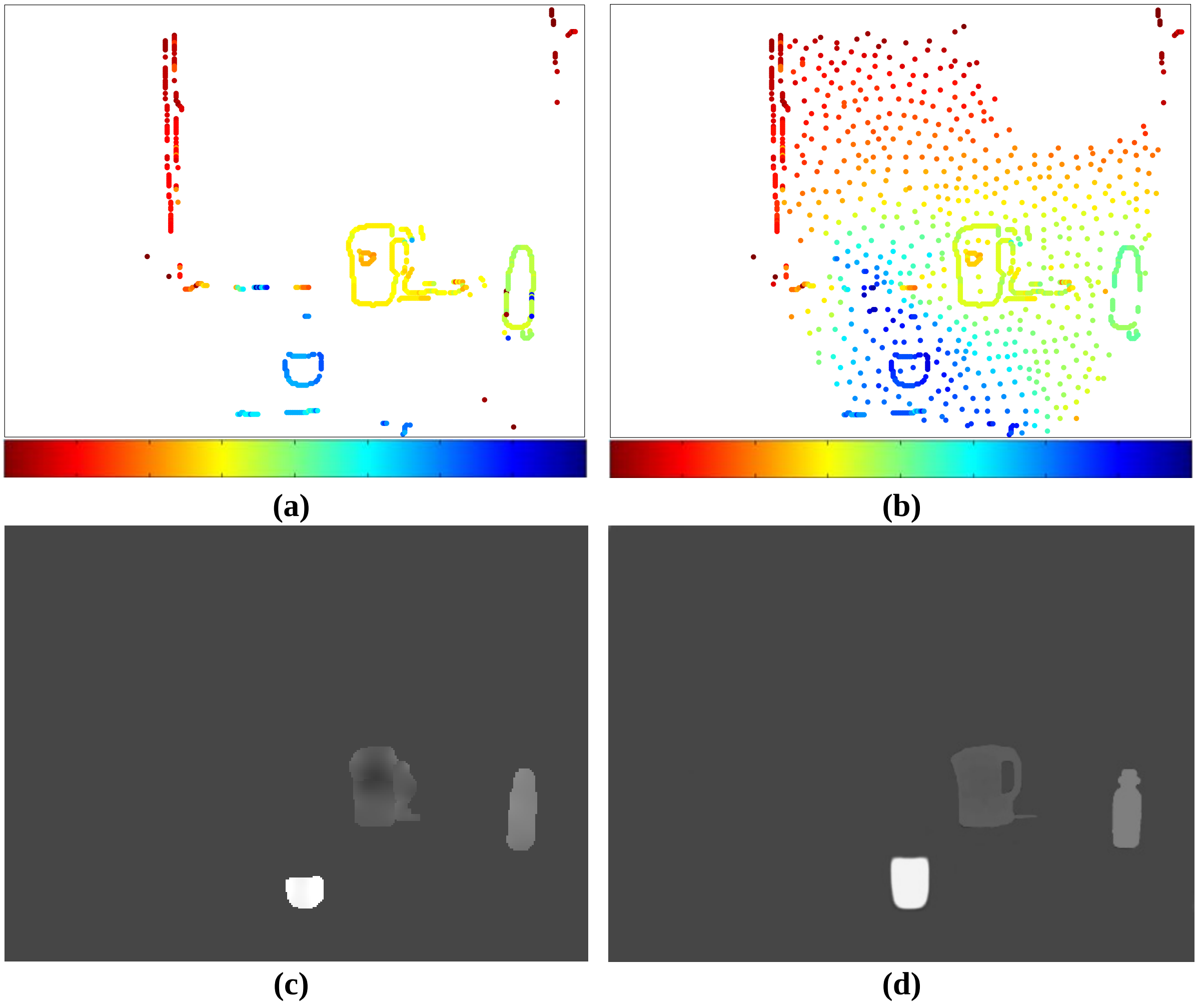
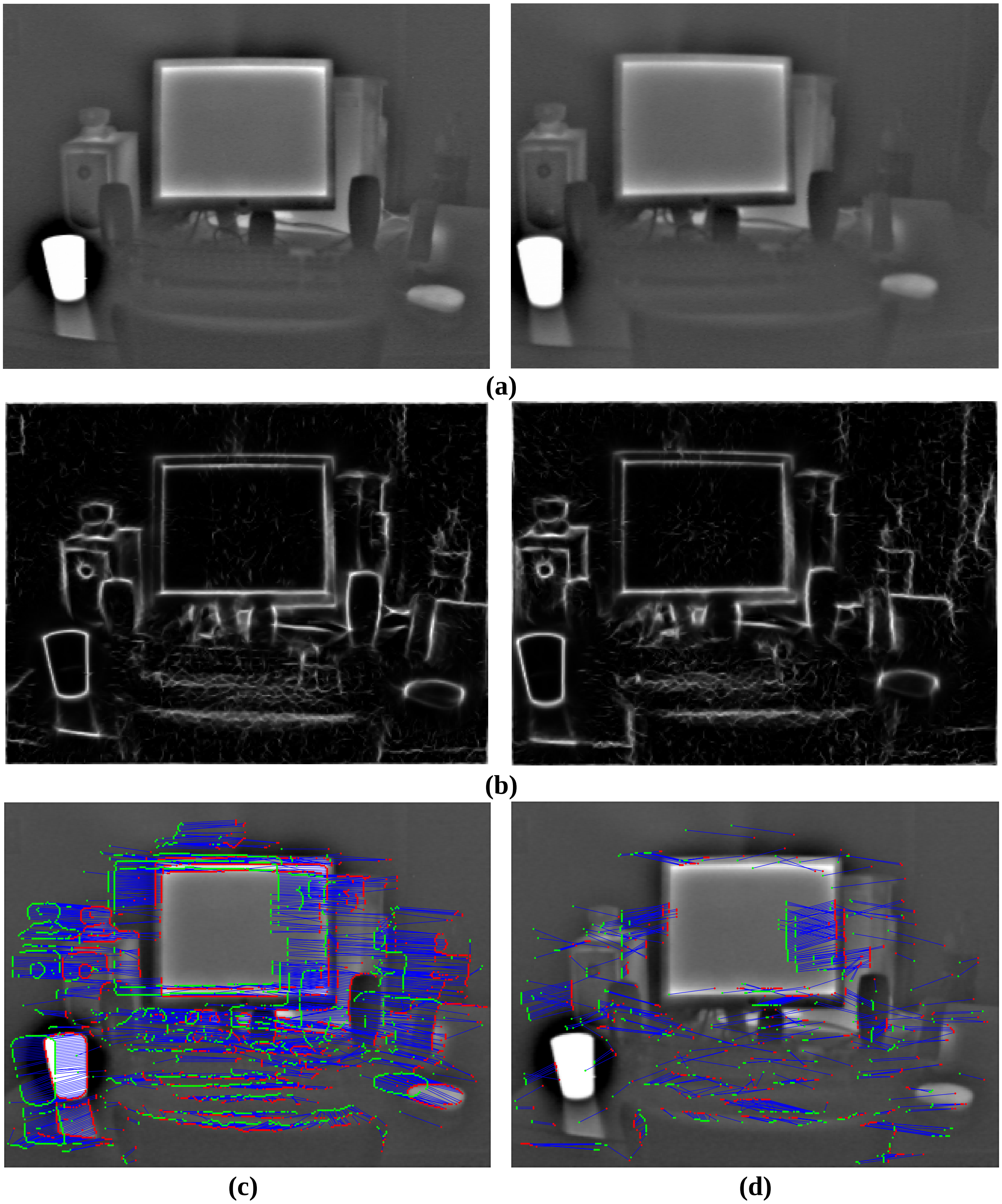


Figure 5.15: (a) sparse disparity map; (b) refined sparse disparity map after Delaunay triangulation; (c) produced dense disparity map; (d) ground truth.


(a)

(b)

(c)

(d)

Figure 5.16: (a) shows the original IR stereo pair; (b) the phase congruency edge maps; (c) displays the left image overlaid with the inliers of the left image (red points) and of the right image (green points), resulted from the matching refinement, along with the correspondences (with blue lines); (d) displays the detected outliers.

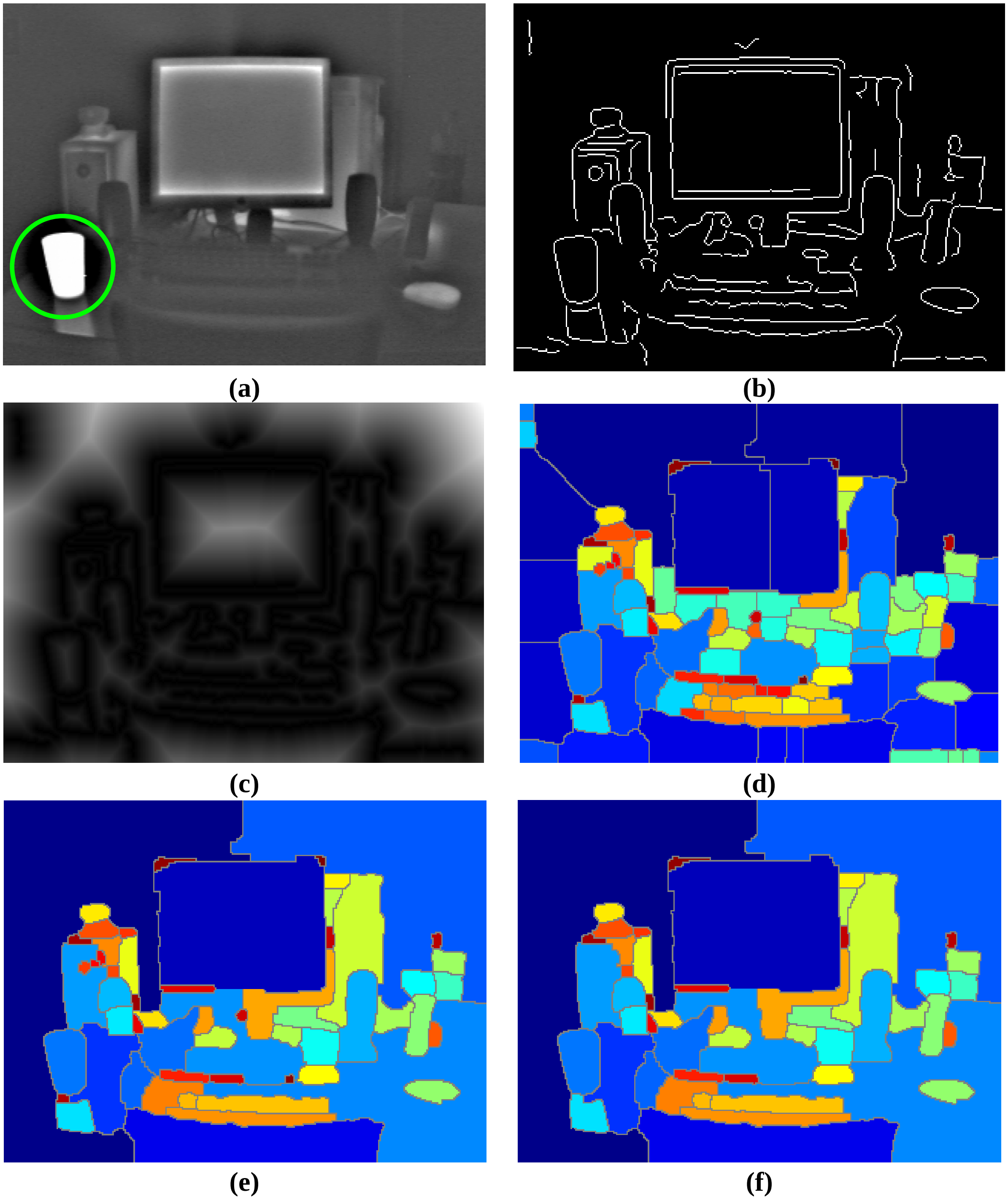


Figure 5.17: (a) left image of an IR stereo pair (reference image); (b) binary edge map; (c) distance transform of the binary map; (d) initial segmentation result; (e) segmentation result after removing irrelevant watershed lines; (f) segmentation result after removing irrelevant watershed lines and removing/merging small regions.

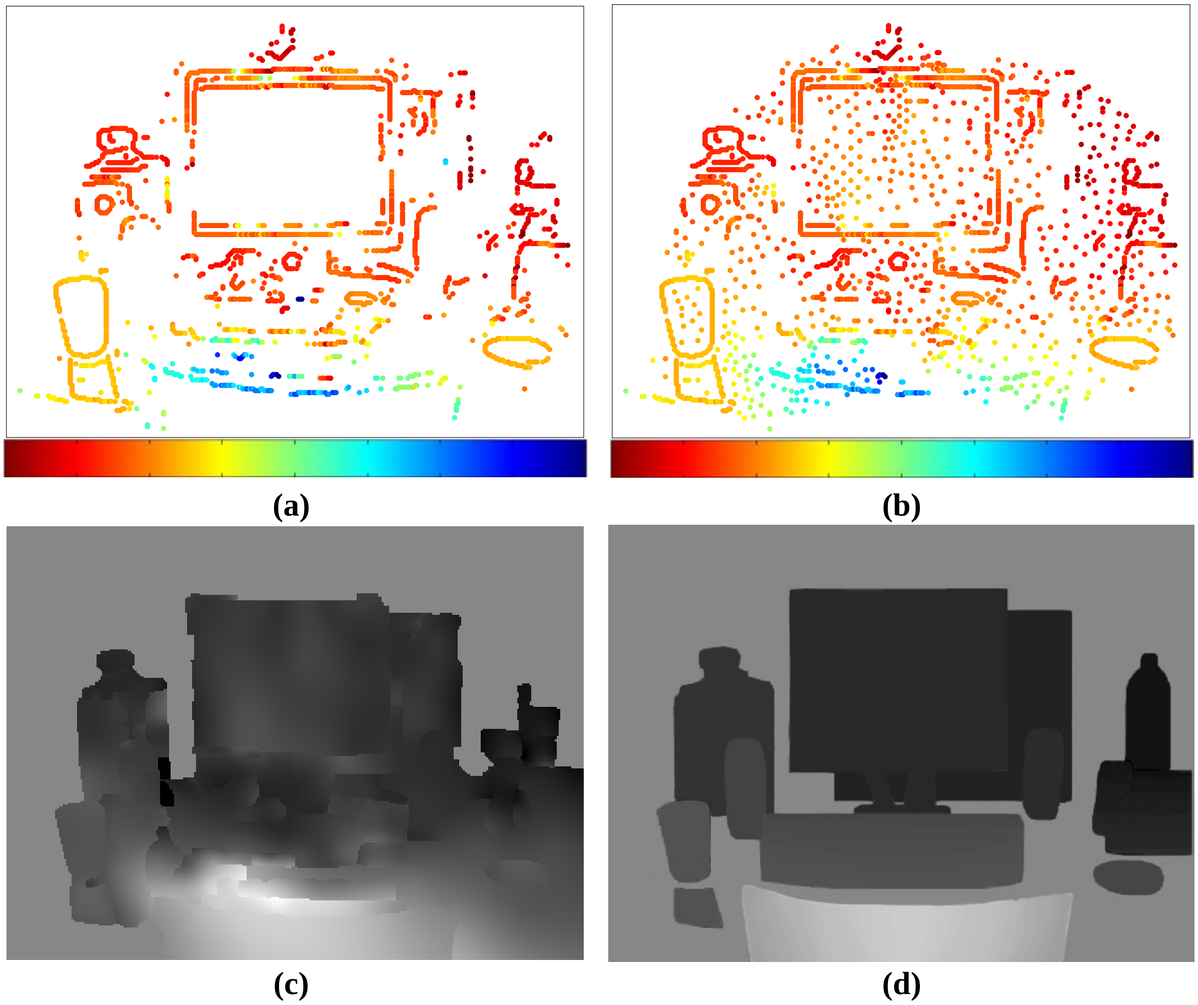
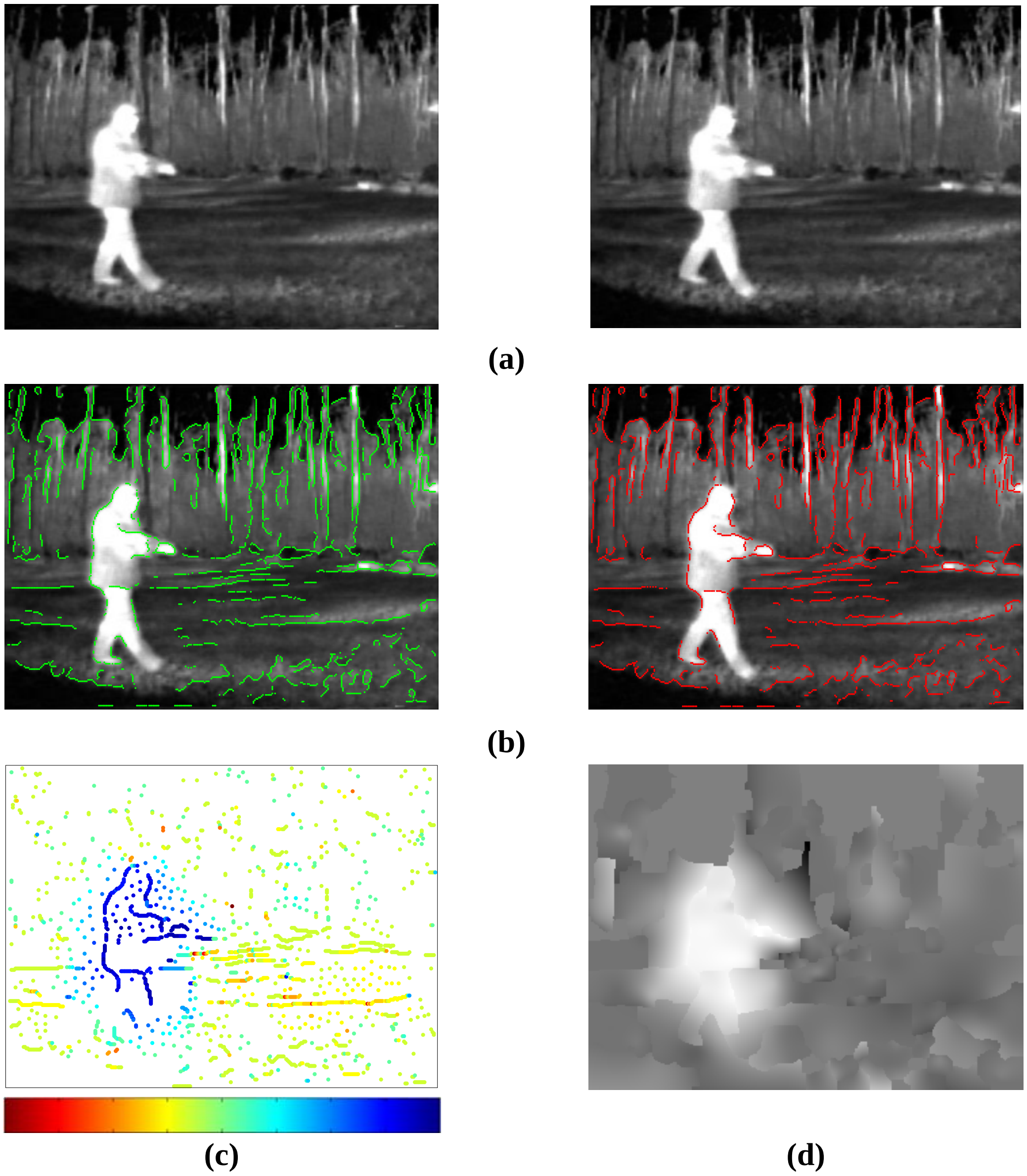


Figure 5.18: (a) sparse disparity map; (b) refined sparse disparity map after Delaunay triangulation; (c) produced dense disparity map; (d) ground truth.



(a)

(b)

(c)

(d)

Figure 5.19: An example of surface reconstruction on an outdoor IR pair. (a) Original IR pair taken from a motion infrared sequence; (b) matched features from (a); (c) sparse disparity map; since the camera is not moving, the disparities of non-moving objects (background) are zero (d) dense disparity map.

Chapter 6

Conclusions and future work

We have presented a novel technique for the computation of semi-dense disparity map from infrared stereo images, and evaluated it on a dataset of stereo IR images from indoor scenes. Our method has shown the ability to compute reliable disparity information in infrared domain for foreground objects in the scenes, without taking into account any prior knowledge about the content of scenes.

6.1 Contributions

In contrast to the previous works ([66] [97]) in the area of computational stereo for infrared images, which showed that the quality (i.e., resolution) of infrared sensors is insufficient for calculating dense depth maps, in this thesis, we have challenged their results and illustrated that a dense depth field may not be attained directly, but perhaps a sparse depth field can be obtained that can be further interpolated to produce a dense/semi-dense depth field.

We have proposed a novel feature-based technique which involves two phases: (i) *feature matching*, i.e., finding a set of corresponding points in the left and the right images to produce a sparse disparity map; and (ii) *reconstruction*, i.e., producing a dense map from a sparse disparity map. For the first phase, we presented a robust IR stereo matching method, which is composed of three main steps. In the first step, a set of stable and tractable feature points from each image is extracted based on the phase congruency model, which contrary to the gradient-based feature detectors, provides features that are invariant to geometric transformations. We obtain the local frequency information for computing the phase congruency via banks of Log-Gabor wavelets at different spatial frequencies and orientations. The wavelet coefficients are further used in describing and matching the extracted features in the second step. Finally, in the last step, the matching results are further analyzed in order to detect and eliminate the outliers, using the epipolar geometrical constraints (Section 3.4.2).

For the second phase, we developed a surface reconstruction technique to densify the sparse disparity map, obtained from the first phase. Our surface reconstruction method consists of three main steps. In the first step, the given sparse disparity map is refined in terms of density rate and measurement distribution. This has been done using triangular and epipolar geometrical constraints (Section 3.4.1). In the second step, the reference image is segmented into homogeneous regions based on its edge map (since the edge features provide the exact, non-blurred locations for the discontinuities), where the disparity can be assumed to vary smoothly inside each region. We achieved this by applying a watershed transformation on the distance transform of the reference image's edge map, along with reasonable image pre- and post- processing to suppress the over-segmentation and obtain a concise region representation (Section 4.3.1). Finally, in the third step, the sparse disparity map is approximated in each region by a surface interpolation technique. Our analysis of several surface fitting methods on synthetic data (Section 4.3.2) prompted us to use thin-plate splines for this step, due to their robustness in the face of high sparsity and noise.

6.2 Future work

With regards to our stereo matching method, one potential direction for future work is to test the efficiency of using different combinations of feature detectors, rather than using only one certain detector, to detect more reliable and tractable features and therefore produce a better sparse disparity map in terms of density rate and measurement distribution. It is unclear if a better feature detector exists but due diligence is necessary before any claims are made.

With regards to our surface reconstruction method, future work is planned to apply the adaptive selection of regularization parameter (lambda) for smoothing surface patches, by using a validation method (e.g., [30]). Using a more sophisticated surface model (rather than thin-plate splines) in the surface fitting process also remains a topic for future work.

Another direction of our future work is to fuse the depth results from infrared cameras (obtained by our method) with other results from visible cameras. Thermal video cameras detect relative differences in the amount of thermal energy emitted/reflected from objects in the scene, making them independent of illumination, and more effective than color cameras under poor lighting conditions. Color sensors on the other hand are oblivious to temperature differences in the scene, and are typically more effective than thermal cameras when objects are at “thermal crossover”, provided that the scene is well illuminated and the objects have color signatures different from the background. Therefore, developing a method which relies on two complementary bands of the electromagnetic spectrum, infrared (thermal) and visible, can perform better than one which relies on an individual band. Preliminary results in Chapter 5 indicate either our method already can do this; or with the use of appropriate detector and representation this should be possible.

Finally, we would like to test our method with real-world applications, including pedestrian detection and tracking for surveillance systems, passenger pose estimation for airbag safety system, robot obstacle detection in dark environment, etc.

Appendix A

Delaunay triangulation

The optimal triangulation of a set of points is one that maximizes the minimum angle in each triangle, producing a set of triangles that are as equilateral as possible [25]. One of the extensively used optimal triangulations is Delaunay triangulation technique, which we chose to use in our method. The Delaunay triangulation for a set P of points in the plane is the triangulation $DT(P)$, defined as follows:

1. Three points $p_i, p_j, p_k \in P$ are vertices in the same face of the Delaunay triangulation *iff* the circle through p_i, p_j, p_k contains no other points. This circle is known as the *circumcircle* of the triangle defined by (p_i, p_j, p_k).
2. Two points $p_i, p_j \in P$ form an edge in the Delaunay triangulation if there is a circle that contains two points p_i, p_j on its boundary and does not contain any other point.

Consequently, the circumcircles of all triangles in $DT(P)$ will contain exactly three points in P on their boundaries if and only if no more than three points in P are co-circular. Delaunay triangulation is usually computed from Voronoi diagram. The relation between the Delaunay triangulation $DT(P)$ and Voronoi diagram $V(P)$ can be described using a concept known as the *dual graph*. The dual graph of a planar graph G has a node for each of the face in G and an arc joining two nodes if their corresponding faces share a common edge. $DT(P)$ is the "straight line" dual graph of $V(P)$. It follows that every edge in $V(P)$ has a corresponding edge in $DT(P)$ and every cell in $V(P)$ has a corresponding point in $DT(P)$ (see Figure A). Algorithms for fast computation of Delaunay triangulation are available in [36].

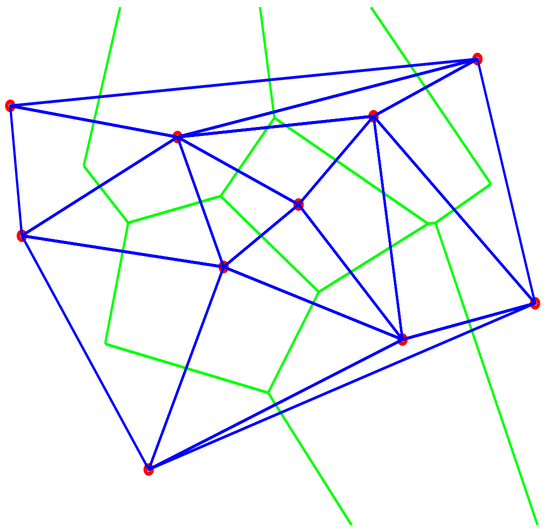


Figure A: The Voronoi diagram and Delaunay triangulation for a set of sample points.

Bibliography

[1] http://vision.middlebury.edu/stereo/eval/.

[2] OTCBVS Benchmark Dataset Collection: http://www.cse.ohio-state.edu/otcbvs-bench/.

[3] F. Arnell, *Vision-based Pedestrian Detection System for use in Smart Cars*, Master's thesis, 2005.

[4] M. Bertozzi, A. Broggi, A. Lasagni and M. D. Rose, "Infrared Stereo Vision-based Pedestrian Detection," In *Procs. of the 2005 IEEE Intelligent Vehicles Symposium*, June 2005.

[5] M. Bertozzi, A. Broggi, A. Fascioli, T. Graf, and M.-M. Meinecke, "Pedestrian Detection for Driver Assistance Using Multiresolution Infrared Vision," *IEEE Trans. on Vehicular Technology*, vol. 53, no. 6, pp. 1666-1678, 2004.

[6] M. Bertozzi, A. Broggi, M. Felisa, G. Vezzoni, and M. D. Rose, "Low-level Pedestrian Detection by means of Visible and Far Infrared Tetra-vision," In *Procs. IEEE Intelligent Vehicles Symposium 2006*, pages 231-236, June 2006.

[7] S. Beucher and F. Meyer, "The morphological approach to segmentation: the watershed transformation," In *Mathematical Morphology in Image Processing*, E. R. Dougherty, Ed. Marcel Dekker, New York, ch. 12, pp. 433-481, 1993.

[8] A. Blake, "The least disturbance principle and weak constraints," *Pattern Recognition Letters,* vol. 1, pp. 393-399, 1983.

[9] A. Blake and A. Zisserman, *Visual Reconstruction*, MIT Press, Cambridge, MA, 1987.

[10] R.M. Bolle, , B.C. Vemuri, "On Three Dimensional Surface Reconstruction Methods," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, pp. 1-14, 1991.

[11] P. Buddharaju, I. Pavlidis, and I. Kakadiaris, "Face recognition in the thermal infrared spectrum," In *Proceedings of the Joint IEEE Workshop on Object Tracking and Classification Beyond the Visible Spectrum*, Washington D.C., 2004.

[12] J. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, Nov. 1986.

[13] W. Carter, "The advantage of single lens stereopsis," in *Proceedings of the SPIE Conference on Stereoscopic Displays and Applications III* , vol. 1669, pp. 204-214, 1992.

[14] R. A. Crone, "The history of stereoscopy," in *Journal of Documenta Ophthalmologica*, vol. 81, no. 1, pp. 1-16, 1992.

[15] R. Cutler, "Face recognition using infrared images and eigenfaces," Tech. Rep., University of Maryland, April 1996. Available at http://www.cs.umd.edu/rgc/pub/ireigenface.pdf

[16] J. W. Davis, V. Sharma, "Robust Background-Subtraction for Person Detection in Thermal Imagery," *IEEE Workshop on Object Tracking and Classification Beyond the Visible Spectrum, CVPR*, 2004.

[17] J. W. Davis, V. Sharma, "Fusion-Based Background-Subtraction using Contour Saliency," In *Procs. of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cvpr'05)*, Workshop, June 2005.

[18] J. W. Davis, and V. Sharma, "Background-Subtraction in Thermal Imagery Using Contour Saliency," *International Journal of Computer Vision*, vol. 71, no. 2, pp. 161-181, 2007.

[19] U. R. Dhond and J. K. Aggarwal, "Structure from stereo-a review," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 19, no. 6, pp. 1489-1510, 1989.

[20] R. G. Driggers, E. L. Jacobs, R. H. Vollmerhausen, B. O'Kane, M. Self, S. Moyer, J. G. Hixson, and G. Page, K. Krapels, D. Dixon, R. Kistner and J. Mazz, "Current infrared target acquisition approach for military sensor design and wargaming," in *Proceedings of SPIE, the International Society for Optical Engineering*, vol. 6207, pp. 620709, 2006.

[21] R. D. Eastman and A. M. Waxman, "Using disparity functionals for stereo correspondence and surface reconstruction," *Computer Vision, Graphics, Image Processing,* vol. 39, 1987.

[22] O. D. Faugeras, M. Herbert, and E. Pauchon, "Segmentation of planar and quadratic patches from range data," in *Proc. IEEE Conf. Pattern Recognition and Image Processing*, 1983.

[23] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *Journal of the Optical Society of America A*, vol. 4, no. 12, pp. 2379-2394, Dec. 1987.

[24] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography," *Commun. Assoc. Comp. Mach.*, vol. 24, no. 6, pp. 381-395, 1981.

[25] J. Fisher, "Visualizing the Connection among Convex Hull, Voronoi Diagram and Delaunay Triangulation," in *the 37th Annual Midwest Instruction and Computing Symposium*, University of Minnesota, Morris, 2004.

[26] X. Fu, Y. Li, R. Harrison, and S. Belkasim, "Content-based image retrieval using Gabor-Zernike features," In *Proceedings of the 18th international Conference on Pattern Recognition (ICPR)*, vol. 2, pp. 417-420. August 20-24, 2006.

[27] P. Fua, “A Parallel Stereo Algorithm that Produces Dense Depth Maps and Preserves Image Features,” *Machine Vision and App.*, vol. 6, no. 1, pp. 35-49, 1993.

[28] S. Geman and D. Geman, “Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images,” *IEEE TPAMI*, vol. 6, no. 6, pp. 721-741, 1984.

[29] O. Ghita, J. Mallon and P.F. Whelan, “Epipolar line extraction using feature matching,” *Proc. Irish Machine Vision and Image Processing Conference 2001*, NUI Maynooth, pp. 87-95, 2001.

[30] G. H. Golub, M. Heath, and G. Wahba, “Generalized cross-validation as a method for choosing a good ridge parameter,” *Technometrics*, vol. 21, pp. 215-- 223, 1979.

[31] R. C. Gonzalez, R. E. Woods, *Digital Image Processing*, 2nd ed. Addison-Wesley Pub. Co., 2002.

[32] E. Goubet, J. Katz and F. Porikli, “Pedestrian Tracking Using Thermal Infrared Imaging,” *SPIE Conference Infrared Technology and Applications XXXII*, vol. 6206, pp. 797-808, 2006.

[33] V. Gouet, P. Montesinos and D. Pele, “A fast matching method for color uncalibrated images using differential invariants,” in *Proceedings of the British Machine Vision Conference*, vol. 1, pp. 367-376, 1998.

[34] W. E. L. Grimson, *From Images to Surfaces: A Computational Study of the Human Early Visual System*, Artificial Intelligence, MIT Press, Cambridge, Massachusetts 1981. Based on the author’s thesis (Ph.D., MIT).

[35] W.E.L. Grimson and T. Pavlidis, “Discontinuity detection for visual surface reconstruction”, *Computer Vision, Graphics, and Image Processing*, vol. 30, pp. 316-330, 1985.

[36] L. Guibas , J. Stolfi, “Primitives for the manipulation of general subdivisions and the computation of Voronoi,” *ACM Transactions on Graphics (TOG)*, vol. 4, no. 2, pp.74-123, 1985.

[37] Y. L. Guilloux and J. Lonnoy, “PAROTO Project: The Benefit of Infrared Imagery for Obstacle Avoidance”, In *Procs. IEEE Intelligent Vehicles Symposium 2002*, June 2002.

[38] M.J. Hannah, “SRI's Baseline Stereo System,” *Proc. of DARPA Image Understanding Workshop*, pp. 149-155, 1985.

[39] K. Haris, S.N. Efstratiadis, N. Maglaveras and A.K. Katsaggelos, “Hybrid image segmentation using watersheds and fast region merging,” *IEEE Transactions on Image Processing*, vol. 7, no. 12, pp. 1684-1699, 1998.

[40] C. Harris, M. Stephens, “A Combined Corner and Edge Detector,” *Alvey Vision Conf.*, pp. 147-151, 1988.

[41] R. Hartley, “In Defense of the Eight-Point Algorithm,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 133-137, 1997.

[42] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision,* Cambridge, U.K.: Cambridge Univ. Press, 2000.

[43] J. Heo, M. Savvides, and B.V.K. Vijayakumar, "Performance Evaluation of Face Recognition using Visual and Thermal Imagery with Advanced Correlation Filters," In *Procs. of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cvpr'05)*, Workshop, vol. 3, 2005.

[44] W. Hoff and N. Ahuja, "Surfaces from stereo Integrating feature matching, disparity estimation and contour detection", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 2, pp. 121-136, Feb. 1989.

[45] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," In *Proceedings of International Conference on Computer Vision*, 2001.

[46] S.G. Kong, J. Heo, B.R. Abidi, J. Paik, M.A Abidi, "Recent advances in visual and infrared face recognition - A review," in *Computer Vision and Image Understanding*, vol. 97, no. 1, pp. 103-135, 2005.

[47] A. Koschan, "A framework for area-based and feature-based stereo vision," *Machine Graphics & Vision*, vol. 2, no. 4, pp. 285-308, 1993.

[48] P. Kovesi, *Invariant Measures of Image Features from Phase Information*, PhD Dissertation, University of Western Australia, 1996.

[49] P. Kovesi, "Image Features from Phase Congruency," *Videre: A Journal of Computer Vision Research*, MIT Press, vol. 1, no. 3, 1999.

[50] P. Kovesi, "Phase Congruency Detects Corners and Edges," In *Proceedings DICTA'03*, pp. 309-318, 2003.

[51] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. Von der Malsburg, R. P. Würtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Trans. on Computers*, vol. 42, no. 3, pp. 300-311, 1993.

[52] A. Leykin and R. Hammoud, "Robust Multi-Pedestrian Tracking in Thermal-Visible Surveillance Videos," In *Procs. of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cvpr'06)*, Workshop, 2006.

[53] S.-S. Lin, "Review: Extending visible bank computer vision techniques to infrared band images," Tech. Rep. MS-CIS-01-04, Computer and Information Science, Univ. of Pennsylvania, Philadelphia, PA, USA, 2001.

[54] X. Liu and K. Fujimura, "Pedestrian detection using stereo night vision," *IEEE Trans. Veh. Technol.*, vol. 53, pp. 1657-1665, 2004.

[55] H. C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, pp.133-135, 1981.

[56] D. Lowe, "Object recognition from local scale-invariant features," In *Proceedings of the7th International Conference on Computer Vision*, pp. 1150-1157, 1999.

[57] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[58] D. Marr and T. Poggio, “A theory of human stereo vision,” Technical Report AI Memo 451, Massachusetts Institute of Technology Artificial Intelligence Laboratory, 1977.

[59] K. Mori, M. Kidode and H. Asada, “An Iterative Prediction and Correction Method for Automatic Stereo comparison,” *Computer Graphics and Image Processing*, vol. 2, pp. 393-401, 1973.

[60] J. Morlet, G. Arens, E. Fourgeau, and D. Giard, “Wave propagation and sampling theory - Part II: Sampling theory and complex waves,” *Geophysics*, vol. 47, no.2, pp. 222-236, 1982.

[61] M. C. Morrone, R. A. Owens, “Feature Detection from Local Energy,” *Pattern recognition letters*, vol. 6, pp. 303-313, 1987.

[62] H. Nanda and L. Davis, “Probabilistic Template Based Pedestrian Detection in Infrared Videos,” in *Procs. IEEE Intelligent Vehicles Symposium 2002*, 2002.

[63] J.A. Noble, “Finding Corners,” *Image and Vision Computing Journal*, vol. 6, no. 2, pp. 121-128, 1988.

[64] Y. Ohta and T. Kanade, “Stereo by Intra- and Inter-Scanline Search Using Dynamic programming,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 7, no. 2, 1985.

[65] V. Okutomi and T. Kanade, “A Locally Adaptive Window for Signal Matching,” *Intl. Journal of Computer Vision,* vol. 7, no. 2, pp. 143-162, 1992.

[66] K. Owens, L. Mathis, “Passive Night Vision Sensor Comparison for Unmanned Ground Vehicle Stereo Vision Navigation,” In *Procs. of the IEEE Workshop on Computer Vision Beyond the Visible*, 1999.

[67] T. Poggio, V. Torre, C. Koch, “Computational vision and regularization theory,” *Nature,* vol. 317, no. 26, pp. 314-319, 1985.

[68] S. J. Pundlik and S. T. Birchfield, “Motion Segmentation at Any Speed,” in *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 427-436, 2006.

[69] B. Remesch and J. M. Cathcart, “Spectral analysis of terrain infrared signatures,” in *Proceedings of SPIE, the International Society for Optical Engineering,* vol. 6217, pp. 62170H, 2006.

[70] S. Roy and I. J. Cox, “A maximum-flow formulation of the N-camera stereo correspondence problem,” in *ICCV*, pp. 492-499, 1998.

[71] A Saaidi, H. Tairi and K. Satori, “Fast Stereo Matching Using Rectification and Correlation Techniques,” in *Proceedings of the 2nd Internatioal Symposium on Communications, control and signal Processing*, 2006.

[72] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” in *International Journal of Computer Vision*, vol. 47, no. (1-3), pp. 7-42, 2002.

[73] C. Schmid, R. Mohr, and C. Bauckhage, “Evaluation of interest point detectors,” in *International Journal of Computer Vision*, vol. 37, no. 2, pp. 151-172, 2000.

[74] J. Shi and C. Tomasi, “Good Features to Track,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 593-600, 1994.

[75] D. Shiwei, Y. Baozong, “A Method for 3D Surface Reconstruction from Range Images”, *International Symposium on Speech, Image Processing and Neural Networks*, pp.13-16, 1994.

[76] D. Socolinsky, L. Wolff, J. Neuheisel, C. Eveland, “Illumination invariant face recognition using thermal infrared imagery,” in *Proceedings CVPR*, Kauai, Dec. 2001.

[77] D. Socolinsky, A. Selinger, “A comparative analysis of face recognition performance with visible and thermal infrared imagery,” in *Proceedings ICPR*, 2002.

[78] D. Socolinsky, A. Selinger, and J. Neuheisel “Face recognition with visible and thermal infrared imagery,” *Computer Vision and Image Understanding*, 2003.

[79] A. Srivastava and X. Liu. “Statistical hypothesis pruning for recognizing faces from infrared images,” *Journal of Image and Vision Computing*, vol. 21, no. 7, pp. 651-661, 2003.

[80] L. Su, C. Luo and F. Zhu, “Obtaining Obstacle Information by an Omni-directional Stereo Vision System,” in *Proceedings of the 2006 IEEE International Conference on Information Acquisition*, pp. 48-52, 2006.

[81] R. Szeliski and R. Zabih, “An experimental comparison of stereo algorithms,” In *International Workshop on Vision Algorithms*, pp. 1-19, 1999.

[82] D. Terzopoulos, “Regularization of inverse visual problems involving discontinuities,” *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 8, no. 4, pp. 413-423, 1986.

[83] D. Terzopoulos, “The computation of visible-surface representations,” *IEEE Transactions on Pattern Analysis and Machine Intelligenc*e, vol. 10, no. 4, pp. 417-438, 1988.

[84] A.N. Tikhonov and V.Y. Arsenin, *Solutions of Ill-posed Problems,* Winston and Sons, Washington, D.C., 1977.

[85] C. Tomasi and T. Kanade, “Detection and tracking of point features,” Technical Report CMU-CS-91-132, Carnegie Mellon University, 1991.

[86] P. H. S. Torr and D. W. Murray, “Statistical detection of independent movement from a moving Camera,” *Image and Vision Computing*, vol. 1, no. 4, pp. 180-187, 1993.

[87] M. M. Trivedi, S. Cheng, E. Childers, S. Krotosky, “Occupant Posture Analysis with Stereo and Thermal Infrared Video: Algorithms and Experimental Evaluation,” *IEEE Transactions on Vehicular Technology, Special Issue on In-Vehicle Vision Systems*, vol. 53, Issue 6, 2004.

[88] T. Tsuji, H. Hattori, M. Watanabe, and N. Nagaoka, “Development of Night-vision System,” *IEEE Trans. on Intelligent Transportation Systems*, vol. 3, pp. 203-209, 2002.

[89] N.M. Vaidya, K.L. Boyer, “Discontinuity preserving surface reconstruction through global optimization,” in *Proceedings of the international IEEE Symposium on Computer Vision*, *ISCV*, pp. 115-120, 1995.

[90] L. Van Gool, T. Moons, and D. Ungureanu, "Affine / photometric invariants for planar intensity patterns," In *ECCV*, pp. 642-651, 1996.

[91] S. Venkatesh and R. A. Owens, "An energy feature detection scheme," In *The International Conference on Image Processing*, pp. 553-557, 1989.

[92] M. Vollmer, S. Henke, D. Kartadt, K. Mollmann, F. Pinno, "Identification and Suppression of Thermal Reflections in Infrared Imaging," InfraMation, 2004.

[93] L. Wiskott, J. Fellous, N. Kruger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. on PAMI*, vol. 19, no. 7, pp. 775-780, 1997.

[94] J. Weickert, "Efficient image segmentation using partial differential equations and morphology," *Pattern Recognition*, vol. 34, no. 9, pp. 1813-1824, 2001.

[95] J. Xin, Z. Yan, **"**The Application of Watershed Algorithm in the Initial Contour Selection of Mumford-Shah Model," in *Proceedings of the 2005 IEEE Engineering in Medicine and Biology Society,* pp. 1773-1776, 2005.

[96] F. Xu and K. Fujimura, "Pedestrian Detection and Tracking with Night Vision," in *Procs. IEEE Intelligent Vehicles Symposium 2002*, June 2002.

[97] J. Zelek, M. Holbein, K. Hajebi and D. Asmar, "IR depth from stereo and context detection for autonomous navigation", in *SPIE Defense and Security Symposium: Infrared Imaging Systems: Design, Analysis, Modeling & Testing XV1*, 2005.

[98] Y. Zhang, C. Kambhamettu, "Stereo matching with segmentation-based cooperation", *European Conference on Computer Vision*, pp. 556-571, 2002.

[99] C. L. Zitnick and T. Kanade, "A cooperative algorithm for stereo matching and occlusion detection," *IEEE TPAMI*, vol. 22, no. 7, pp. 675-684, 2000.