Acquisition and Processing of ToF and Stereo Data

Carlo Dal Mutto

Ph.D. School on Information Engineering
University of Padova

Advisor: Prof. G. M. Cortelazzo

January 13, 2013

Abstract

Providing a computer the capability to estimate the three-dimensional geometry of a scene is a fundamental problem in computer vision. A classical systems that has been adopted for solving this problem is the so-called stereo vision system (stereo system). Such a system is constituted by a couple of cameras and it exploits the principle of triangulation in order to provide an estimate of the framed scene. In the last ten years, new devices based on the time-of-flight principle have been proposed in order to solve the same problem, *i.e.*, matricial Time-of-Flight range cameras (ToF cameras).

This thesis focuses on the analysis of the two systems (ToF and stereo cameras) from a theoretical and an experimental point of view. ToF cameras are introduced in Chapter 2 and stereo systems in Chapter 3. In particular, for the case of the ToF cameras, a new formal model that describes the acquisition process is derived and presented. In order to understand strengths and weaknesses of such different systems, a comparison methodology is introduced and explained in Chapter 4. From the analysis of ToF cameras and stereo systems it is possible to understand the complementarity of the two systems and it is intuitive to figure that a synergic fusion of their data might provide an improvement in the quality of the measurements preformed by the two devices. In Chapter 5 a method for fusing ToF and stereo data based on a probability approach is presented. In Chapter 6 a method that exploits color and three-dimensional geometry information for solving the classical problem of scene segmentation is explained.

Sommario

Fornire ai calcolatori la capacità di stimare la geometria tridimensionale di una scena è una delle sfide fondamentali nell'ambito della visione artificiale. Il classico approccio utilizzato per la risoluzione di tale problema prevede l'utilizzo di sistemi di visione stereoscopica. Tali sistemi sono costituiti da due telecamere. Il loro funzionamento si basa sul principio di triangolazione per stimare la configurazione geometrica di una scena. Nell'ultimo decennio, nuovi dispositivi basati sul principio del tempo di volo sono stati proposti allo scopo di risolvere il medesimo problema. Tali dispositivi sono chiamati sensori di profondità matriciali a tempo di volo.

Questa tesi si sviluppa attorno all'analisi dei suddetti sistemi da un punto di vista teorico e sperimentale. I sensori a tempo di volo vengono descritti nel Capitolo 2, mentre i sistemi stereo nel Capitolo 3. In particolare viene introdotto un nuovo modello che descrive formalmente il processo di acquisizione dei sensori a tempo di volo. Nel Capitolo 4 viene descritta una metodologia per confrontare i due diversi sistemi. Da questa analisi emerge chiaramente la complementarietà dei due sistemi. Questo permette di intuire come una fusione dei loro dati renda possibile un miglioramento della stima geometrica. Nel Capitolo 5 viene descritto un metodo che consente di fondere i dati del sistema stereo e del sensore a tempo di volo. Nel Capitolo 6 viene sviluppato un metodo per sfruttare l'informazione sul colore e sulla geometria di una scena per risolvere il classico problema di segmentazione della scena.

A Chiara, che mi ha sempre sostenuto.

Acknowledgements

First of all I would like to thank Prof. Guido M. Cortelazzo for the wise and stimulating mentorship during my Ph.D. studies.

I am thankful to Chiara, which always considers my “dangerous ideas” that leads to sudden changes in our life. She has been of great support during this period.

I am grateful to my parents Marina and Umberto for the continuous stimulation in pursuing new ideas and adventures, to my brother Marco and to Linda for being always there, and to my family for the generous support.

During these years of studies, the collaboration with several colleagues has been a great fortune for me. I would like to thank Pietro Zanuttigh and Stefano Mattoccia for the availability and the support they gave me.

I am in debt with Prof. Carlo Tomasi, which was extremely kind and illuminating during my time at Duke University.

I want to thank also Arrigo Benedetti and Roberto Manduchi, wise collaborators; Abbas Rafii and Nazim Kareemi, who have strongly believed in me; and Michele Stoppa and Giuliano Pasqualotto for being great fellows in my adventures.

I am grateful to the various students that helped me during my research and to all my friends and collegues.

This work has been generously supported by the “Fondazione Ing. Aldo Gini”, to which I owe my thank.

Contents

1 Introduction **1**

2 Matricial Time-of-Flight (ToF) range cameras **5**

2.1 CW ToF sensors: operation principles 6

2.2 CW ToF sensors: practical implementation issues 8

2.2.1 Phase wrapping 8

2.2.2 Harmonic distortion 8

2.2.3 Photon-shot noise 9

2.2.4 Other noise sources 11

2.2.5 Saturation and motion blur 12

2.3 Matricial ToF cameras 12

2.3.1 Matricial ToF sensors 13

2.3.2 ToF Camera imaging characteristics 15

2.3.3 Practical imaging issues 18

3 Stereo vision systems **23**

3.1 Stereo vision algorithms 26

3.2 Local stereo algorithms 27

3.3 Global stereo algorithms 28

3.4 Semi-global stereo algorithms 30

4 Comparison of ToF cameras and stereo systems in terms of metrological quantities **31**

4.1 Accuracy, precision and resolution of ToF cameras and stereo systems . 32

4.1.1 Accuracy 32

4.1.2 Precision 34

4.1.3 Resolution 36

4.2 Experimental comparison 38

5 Fusion of tof and stereo data: probabilistic approach 43

5.1 Related Work 43

5.2 Depth Estimation from Multiple Devices Measurements 46

5.2.1 Depth Estimation from ToF and Stereo Data 48

5.3 ToF Likelihood 51

5.4 Stereo Likelihood 56

5.5 Scene Depth Prior 58

5.6 Building Posterior 62

5.7 Loopy-Belief-Propagation optimization 63

5.8 Experimental Results 64

6 Scene segmentation from 3D and color data 79

6.1 Introduction 79

6.2 Joint representation of geometry and color information 81

6.3 Segmentation by means of spectral clustering and Ny ström method . . 85

6.4 Automatic weighting of color and depth information 87

6.5 Segmentation of stereo image pairs 93

6.6 Experimental Results 95

6.6.1 Results on the trinocular system data 97

6.6.2 Results on stereo vision data 98

6.6.3 Results on Kinect data 103

6.6.4 Results from Photosynth data 108

6.7 Comparison of the considered imaging systems for scene segmentation purposes 108

7 Conclusion 113

Bibliography 115

1

Introduction

Three-dimensional data acquisition is the task of acquiring information about the geometrical configuration of a particular portion of the space. It is a standard sensing problem which can be tackled by means of different techniques, depending on the characteristic of the considered space. In this thesis the focus is on dynamic scenes characterized by maximum distances with respect to the cameras in the order of few meters (*e.g.*, $500 - 5000[mm]$). Classical approaches for acquiring three-dimensional information of such scenes presume the exploitation of stereo vision systems (or simply stereo systems).

Stereo vision techniques have been considered in the last few decades, they have matured, but they still are not able to completely solve the problem [86]. Some of the advantages of stereo techniques are the capability of delivering potentially cheap, high resolution and precision three-dimensional information of the acquired scene. Also they are suited for different illumination scenarios since the stereo acquisition system is constituted by standard cameras, the components of which can be selected according to the considered scenario. Among all the problems typical of these systems, the so-called *aperture problem*, *i.e.*, the inability to deal with textureless scenes, is one of the most crucial.

More recently new types of systems that aim at solving the same problem have been introduced, *i.e.*, Time-of-Flight range cameras (simply ToF cameras) [45, 88]. These cameras are active systems that irradiate at InfraRed (IR) wavelength, and collect the signal reflected back by the scene. Since they are active systems, they are characterized by a greater power consumption with respect to stereo systems, but they are generally able to acquire three-dimensional geometry information of the scene more robustly. ToF cameras are generally characterized by higher accuracy and precision, but lower resolution with respect to stereo systems.

The complementarity of the characteristics of the stereo systems and ToF cameras in terms of three-dimensional information acquisition suggests that a synergic fusion of the data acquired by the two subsystems can be performed in order to obtain a superior quality of acquired three-dimensional geometry information. One of the main focuses of this thesis is the investigation of how to take advantage of the best characteristics of the two systems in order to fuse their data.

An important characteristic of this acquisition setup is the possibility of acquiring both color and three-dimensional geometry of the framed scene. Such multi-modal information can be widely exploited in a number of applications, such as scene segmentation, object recognition, people detection, body tracking, hand-gestures recognition and many more else. Among all applications, the scene segmentation problem is tackled in this thesis. Scene segmentation is the problem of identifying all the different elements in a scene. Historically this problem has been tackled by means of standard images, hence it has usually been called image segmentation. One important contribution of this thesis is the investigation of the improvement that three-dimensional geometry information can provide for the solution of this problem.

This thesis comes at the conclusion of my Ph.D. studies and it reflects the main stream of my investigation in these three exciting years. As the title suggests, the principal theme is the acquisition and processing of ToF and stereo data.

The first steps that were done during this period are towards the analysis and comprehension on how ToF cameras and stereo systems work and what is the quality of their acquired data. While for stereo systems the prior art is extremely vast, the literature regarding ToF is still in a rapid evolution. In order to provide a solid foundation for ToF and stereo data fusion, data acquisition models for ToF have been investigated and published into [41, 42, 45]. In particular [41] was awarded with the *best paper award* prize at the GTTI meeting 2010. The most advanced model for ToF data acquisition is the one presented in [45], which is also presented and further developed in this thesis.

Once established a basilar knowledge of ToF cameras and the literature of stereo systems analyzed, a rigorous comparison of stereo and ToF performances has been done in order to assess the complementarity if the two systems. In fact, while it is immediate to see how the nature of the two systems is different, a rigorous comparison in terms of classical metrological quantities such as accuracy precision and resolution is necessary in order to validate this concept. Such comparison was proposed in [46] of [88] and showed how ToF and stereo are complementary in terms of the introduced metrological quantities. This analysis provided a good motivation for pursuing the problem.

A preliminary operation that has to be performed in order to allow ToF and stereo data fusion is the setup and the calibration of the trinocular system obtained by composition of a ToF cameras and a stereo system. System calibration had been tackled and a closed-form solution was proposed in [42] and further expanded in [45].

The problem of ToF and stereo fusion was approached with local probabilistic methods [41, 42] based on a Maximum-Likelihood (ML) Bayesian formulation and with local consistency [48] approaches. The method presented is this thesis constitutes a substantial improvement of the maximum likelihood Bayesian formulation of [41, 42]. In fact the considered method is a Maximum-a-Posteriori Markov Random Field (MAP-MRF) Bayesian approach, which exploits the formal model of ToF cameras that has been developed.

Given the data obtained from the trinocular system some applications have been considered. In particular the focus has been on scene segmentation and hand gestures recognition. Concerning scene segmentation, a first approach was introduced in [43] and a more mature version in [44].

Hand gestures have been approached in [39], which was assigned the *best paper award* at the STreaming Day 2011 (annual conference of the STMicroelectronics research group [19]).

Several interesting divagations also occurred, such as a comparison of ToF cameras and the Microsoft Kinect [9] structured light camera [22], an approach to scene segmentation from stereo data [47] and the study of the joint solution of scene segmentation and stereo depth estimation as a unique problem [40].

During the Ph.D I also had the chance to look at 3D acquisition systems from a more general point of view while writing the book "Time-of-Flight Cameras and Microsoft Kinect$^{\mathrm{TM}}$" [45].

In order to preserve readability, not all these topics are exposed in this thesis. In particular, Chapter 2 presents the proposed formal model for ToF data acquisition and error analysis. Such models not only accounts for classical per-pixel acquisition models, but also consider the finite size of pixels in ToF cameras in order to obtain a more comprehensive analysis. Chapter 3 analyzes stereo vision systems by reviewing the literature from the perspective of the fusion problem. Chapter 4 defines the classical metrological concepts of accuracy, precision and resolution for scene depth acquisition systems and provides an analysis of such quantities for ToF cameras and stereo systems. Chapter 5 presents a Maximum-a-Posteriori Markov Random Field (MAP-MRF) Bayesian approach to ToF and stereo data fusion. Such MAP-MRF approach exploits the proposed ToF model and the global optimization of the MAP problem is performed with an extension of Loopy Belief Propagation to problems characterized by site-dependent domain. Chapter 6 introduces a solution of the scene segmentation problem that synergically accounts for both geometry and color information. Finally Chapter 7 draws the conclusions [1].

[1] Some of the chapters' contents are taken from previously published material. In particular, Chapter 2 is from [45], Chapters and 3 and 4 from [46] and Chapter 6 from [44].

2

Matricial Time-of-Flight (ToF) range cameras

Matricial Time-of-Flight range cameras (simply *ToF cameras*) are active sensors capable to acquire the three-dimensional geometry of the framed scene at video rate (up to $50\,[fps]$). Commercial products are currently available from independent manufacturers, such as MESA Imaging [8] (Figure 2.1), PMD Technologies [16] and SoftKinetic [18]. Microsoft [11] is another major actor in the ToF camera technology arena since at the end of 2010 it acquired Canesta, a U.S. ToF camera manufacturer. Other companies (*e.g.*, Panasonic [15] and IEE [6]) and research institutions (*e.g.*, CSEM [1] and Fondazione Bruno Kessler [4]) are also working on ToF cameras.

This chapter examines continuous wave ToF technology, which is the technological basis of all the current commercial products. Section 2.1 presents the operating principles of such technology and Section 2.2 the practical issues at the basis of its performance limits and noise characteristics. The characteristics of ToF cameras, *i.e.*,

Figure 2.1: Example of commercial ToF camera: MESA SR4000$^{\text{TM}}$.

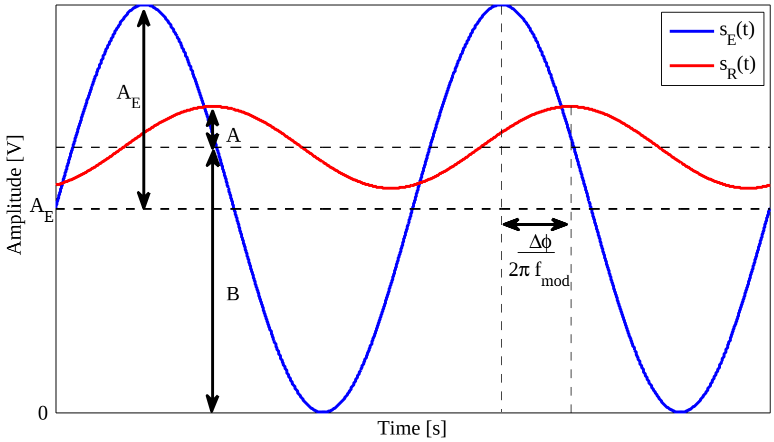


Figure 2.2: Example of emitted signal $s_E(t)$ (in blue) and received signal $s_R(t)$ (in red).

of the imaging system supporting ToF sensors, are considered in Section 2.3.

2.1 CW ToF sensors: operation principles

Continuous wave ToF cameras send towards the scene an infra-red (IR) optical signal $s_E(t)$ of amplitude A_E modulated by a sinusoid of frequency f_{mod}, namely

$$s_E(t) = A_E[1 + \sin(2\pi f_{mod} t)] \tag{2.1}$$

Signal $s_E(t)$ is reflected back by the scene surface and travels back towards a receiver co-positioned with the emitter.

The signal reaching the receiver, because of the energy absorption generally associated to the reflection, because of free-path propagation attenuation (proportional to the square of the distance) and because of the non-instantaneous propagation of IR optical signals leading to a phase delay $\Delta\phi$, can be written as

$$s_R(t) = A_R[1 + \sin(2\pi f_{mod} t + \Delta\phi)] + B_R \tag{2.2}$$

where A_R is the attenuated amplitude of the received signal and B_R is the interfering radiation at the IR wavelength of the emitted signal reaching the receiver. Figure 2.2 shows an example of emitted and received signals. Quantity A_R (from now denoted by A) is called *amplitude*, since it is the amplitude of the useful signal. Quantity $A_R + B_R$

(from now denoted by B) is called *intensity* or *offset*, and it is the average[1] of the received signal (with a component A_R due to the modulation carrier and an interference component B_R due to background illumination). According to this notation, Equation (2.2) can be rewritten as

$$s_R(t) = Asin(2\pi f_{mod} t + \Delta\phi) + B \tag{2.3}$$

The unknowns of Equation (2.3) are A, B and $\Delta\phi$, where A and B as IR radiation amplitudes are measured in volt $[V]$ and $\Delta\phi$ as a phase value is a pure number. The most important unknown is $\Delta\phi$, since CW ToF cameras infer distance ρ from $\Delta\phi$ and it can be computed as

$$\Delta\phi = 2\pi f_{mod}\tau = 2\pi f_{mod}\frac{2\rho}{c} \tag{2.4}$$

or equivalently

$$\rho = \frac{c}{4\pi f_{mod}}\Delta\phi \tag{2.5}$$

Unknowns A and B as it will be seen are important for SNR considerations.

In order to estimate the unknowns A, B and $\Delta\phi$, the receiver samples $s_R(t)$ at least 4 times per period of the modulating signal [73]. For instance, if the modulation frequency is $30[MHz]$, the received signal must be sampled at least at $120[MHz]$. Assuming a sampling frequency $F_S = 4f_{mod}$, given the 4 samples per period $s_R^0 = s_R(t = 0)$, $s_R^1 = s_R(t = 1/F_S)$, $s_R^2 = s_R(t = 2/F_S)$ and $s_R^3 = s_R(t = 3/F_S)$, the receiver estimates values $\hat{A}$,$\hat{B}$ and $\widehat{\Delta\phi}$ as

$$(\hat{A}, \hat{B}, \widehat{\Delta\phi}) = \arg \min_{A,B,\Delta\phi} \sum_{n=0}^{3} \{s_R^n - [Asin(\frac{\pi}{2}n + \Delta\phi) + B]\}^2 \tag{2.6}$$

As described in [32] and [80], after some algebraic manipulations from (2.6) one obtains

$$\hat{A} = \frac{\sqrt{\left(s_R^0 - s_R^2\right)^2 + \left(s_R^1 - s_R^3\right)^2}}{2} \tag{2.7}$$

$$\hat{B} = \frac{s_R^0 + s_R^1 + s_R^2 + s_R^3}{4} \tag{2.8}$$

$$\widehat{\Delta\phi} = \arctan 2\left(s_R^0 - s_R^2, s_R^1 - s_R^3\right) \tag{2.9}$$

[1] It is common to call A and B *amplitude* and *intensity* respectively, even though both A and B are IR radiation amplitudes (measured in $[V]$). A is also the amplitude of the received sinusoidal signal.

The final distance estimate $\hat{\rho}$ can be obtained combining (2.5) and (2.9) as

$$\hat{\rho} = \frac{c}{4\pi f_{mod}} \widehat{\Delta\phi} \tag{2.10}$$

2.2 CW ToF sensors: practical implementation issues

The above derivation highlights the conceptual steps needed to measure the distance ρ of a scene point from a CW ToF sensor, with co-positioned emitter and receiver. In practice a number of non-idealities, such as *phase wrapping*, *harmonic distortion*, *noise sources*, *saturation* and *motion blur*, must be taken into account.

2.2.1 Phase wrapping

The first fundamental limitation of CW ToF sensors comes from the fact that the estimate of $\widehat{\Delta\phi}$ is obtained from an arctangent function, which has codomain $[-\frac{\pi}{2}, \frac{\pi}{2}]$. The estimates of $\widehat{\Delta\phi}$ can only assume values in this interval. Since the physical delays entering the phase shift $\Delta\phi$ of Equation (2.4) can only be positive, it is possible to shift the $\arctan(\cdot)$ codomain to $[0, \pi]$ in order to have a larger interval available for $\widehat{\Delta\phi}$. Moreover, the usage of $\arctan 2(\cdot, \cdot)$ allows to extend the codomain to $[0, 2\pi]$. From Equation (2.10) it is immediate to see that the estimated distances are within range $[0, \frac{c}{2f_{mod}}]$. If for instance $f_{mod} = 30[MHz]$, the interval of measurable distances is $[0 - 5000][mm]$.

Since $\widehat{\Delta\phi}$ is estimated modulo 2π from (2.10) and the distances greater than $\frac{c}{2f_{mod}}$ correspond to $\widehat{\Delta\phi}$ greater than 2π, they are wrongly estimated. In practice the distance returned by (2.10) corresponds to the remainder of the division between the actual $\Delta\phi$ and 2π, multiplied by $\frac{c}{2f_{mod}}$, a well-known phenomenon called *phase wrapping* since it may be ragarded as a periodic wrapping around 2π of phase values $\widehat{\Delta\phi}$. Clearly if f_{mod} increases, the interval of measurable distances becomes smaller, and vice-versa. Possible solutions to overcome phase wrapping include the usage of multiple modulation frequencies or of non-sinusoidal wave-forms (*e.g.*, chirp wave-forms).

2.2.2 Harmonic distortion

The generation of perfect sinusoids with the required frequency is not straightforward. In practice [33], actual sinusoids are obtained as low-pass filtered versions of squared

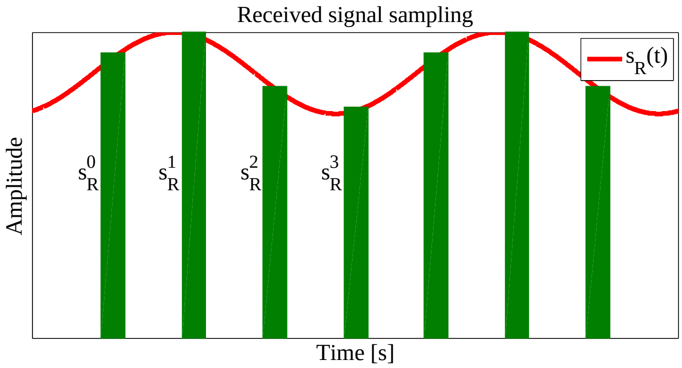


Figure 2.3: Pictorial illustration of non instantaneous sampling of the received signal $s_R(t)$.

wave-forms emitted by LEDs. Moreover, the sampling of the received signal is not ideal, but it takes finite time intervals, as shown in Figure 2.3. The combination of these two factors introduces an harmonic distortion in the estimated phase-shift $\widehat{\Delta\phi}$ and consequently in the estimated distance $\hat{\rho}$. Such harmonic distortion leads to a systematic offset component dependent on the measured distance. A metrological characterization of this harmonic distortion effect is reported in [67] and [93].

Figure 2.4 shows that the harmonic distortion offset exhibits a kind of oscillatory behavior which can be up to some tens of centimeters, clearly reducing the accuracy of distance measurements. This systematic offset can be fixed by a look-up-table (LUT) correction [45].

2.2.3 Photon-shot noise

Because of the light-collecting nature of the receiver, the acquired samples s_R^0, s_R^1, s_R^2 and s_R^3 are affected by photon-shot noise, due to dark electron current and photon-generated electron current, as reported in [32]. Dark electron current can be reduced by lowering the sensor temperature or by technological improvements. Photon-generated electron current, due to light-collection, cannot be completely eliminated. Photon-shot noise is statistically characterized by a Poisson distribution. Since $\hat{A}$, $\hat{B}$, $\widehat{\Delta\phi}$ and $\hat{\rho}$ are computed directly from the corrupted samples s_R^0, s_R^1, s_R^2 and s_R^3, their noise distribution can be computed by propagating the Poisson distribution through Equations (2.7-2.10). A detailed analysis of error and noise propagations can be found in [80].

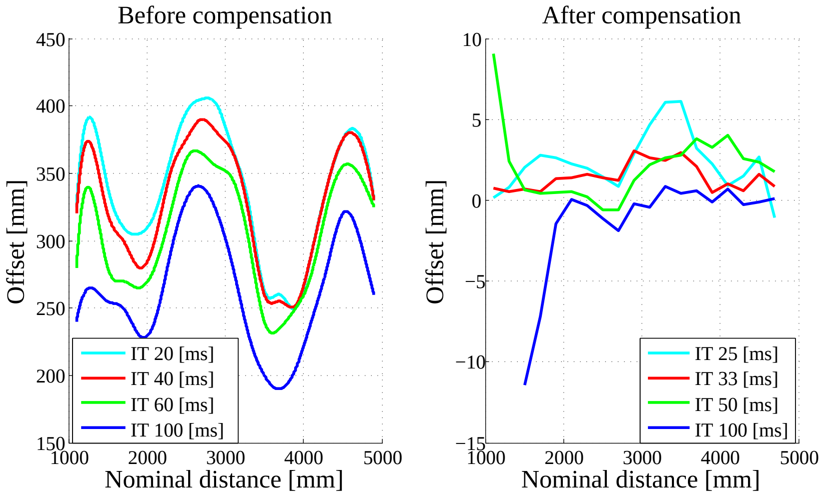


Figure 2.4: Left: systematic distance measurements offset due to harmonic distortion before compensation (from [67]). Right: systematic distance measurements offset after compensation (courtesy of MESA Imaging).

The probability density function of the noise affecting estimate $\hat{\rho}$ according to [32] and [80] can be approximated by a Gaussian. However, the model of [80] provides implicit information about the mean which is a function of both A and B, and contributes to the distance measurement offset. For calibration purposes the non-zero mean effect can be included in the harmonic distortion with standard deviation (and mean)

$$\sigma_\rho = \frac{c}{4\pi f_{mod}\sqrt{2}} \frac{\sqrt{B}}{A} \tag{2.11}$$

Standard deviation (2.11) determines the precision (repeatability) of the distance measurement and it is directly related to f_{mod}, A and B. In particular, if the received signal amplitude A increases, the precision improves. This suggests that the precision improves as the measured distance decreases and the reflectivity of the measured scene point increases.

Equation (2.11) indicates also that as the interference intensity B of the received signal increases, the precision gets worse. This means that the precision improves as the scene background IR illumination decreases. Note that B may increase because of two factors: an increment of the received signal amplitude A or an increment of the background illumination. While in the second case the precision gets worse, in the first case there is an overall precision improvement, given the squared root dependence of B in

(2.11). Finally it is worth to observe that B cannot be zero as it depends on carrier intensity A.

If modulation frequency f_{mod} increases the precision improves. The modulation frequency is an important parameter for ToF sensors, since f_{mod} is also related to phase wrapping and to the maximum measurable distance. In fact, if f_{mod} increases the measurement precision improves, while the maximum measurable distance decreases (and vice-versa). Therefore there is a trade-off between distance precision and range. Since generally f_{mod} is a tunable parameter, it can be adapted to the distance precision and range requirements of the specific application.

2.2.4 Other noise sources

There are several other noise sources affecting the distance measurements of ToF sensors, namely *flicker* and a *kTC noise*. The receiver amplifier introduces a Gaussian-distributed thermal noise component. Since the amplified signal is quantized in order to be digitally treated, quantization introduces another error source, customarily modeled as random noise. Quantization noise can be controlled by the number of used bits and it is typically neglectable with respect to the other noise sources. All the noise sources, except photon-shot noise, may be reduced by adopting high quality components. A comprehensive description of the various ToF noise sources can be found in [32, 33, 73, 80].

Averaging distance measurements over several periods is a classical provision to mitigate the noise effects. If N is the number of periods, the estimated values $\hat{A}$, $\hat{B}$ and $\widehat{\Delta\phi}$ become

$$\hat{A} = \frac{\sqrt{\left(\frac{1}{N}\sum_{n=0}^{N-1} s_R^{4n} - \frac{1}{N}\sum_{n=0}^{N-1} s_R^{4n+2}\right)^2 + \left(\frac{1}{N}\sum_{n=0}^{N-1} s_R^{4n+1} - \frac{1}{N}\sum_{n=0}^{N-1} s_R^{4n+3}\right)^2}}{2} \tag{2.12}$$

$$\hat{B} = \frac{\sum_{n=0}^{N-1} s_R^{4n} + \sum_{n=0}^{N-1} s_R^{4n+1} + \sum_{n=0}^{N-1} s_R^{4n+2} + \sum_{n=0}^{N-1} s_R^{4n+3}}{4N} \tag{2.13}$$

$$\widehat{\Delta\phi} = \arctan 2\left(\frac{1}{N}\sum_{n=0}^{N-1} s_R^{4n} - \frac{1}{N}\sum_{n=0}^{N-1} s_R^{4n+2}, \frac{1}{N}\sum_{n=0}^{N-1} s_R^{4n+1} - \frac{1}{N}\sum_{n=0}^{N-1} s_R^{4n+3}\right) \tag{2.14}$$

where $s_R^{4n} = s_R(4n/F_S)$, $s_R^{4n+1} = s_R((4n+1)/F_S)$, $s_R^{4n+2} = s_R((4n+2)/F_S)$ and $s_R^{4n+3} = s_R((4n+3)/F_S)$.

This provision reduces but does not completely eliminate the noise effects. The averaging intervals used in practice are typically between $1[ms]$ and $100[ms]$. For instance in case of $f_{mod} = 30MHz$, where the modulating sinusoid period is $33.3 \times 10^{-9}[s]$, the averaging intervals concern a number of modulating sinusoid periods from 3×10^4 to 3×10^6. The averaging interval length is generally called *integration time*, and its proper tuning is very important in ToF measurements. Long integration times lead to good ToF distance measurements repeatability.

2.2.5 Saturation and motion blur

Although rather effective against noise, averaging over multiple periods introduces dangerous side effects, such as *saturation* and *motion blur*. Saturation occurs when the received photons quantity exceeds the maximum quantity that the receiver can collect. This phenomenon is particularly noticeable in presence of external IR illumination (*e.g.*, direct solar illumination) or in case of highly reflective objects (*e.g.*, specular surfaces). The longer the integration time, the higher is the quantity of collected photons and the most likely is the possibility of saturation. Specific solutions have been developed in order to avoid saturation, *i.e.*, in-pixel background light suppression and automatic integration time setting [32, 33].

Motion blur is another important phenomenon accompanying time averaging. It is caused, as in the case of standard cameras, by the fact that the imaged objects may move during integration time. Time intervals of the order of $1 - 100[ms]$ make likely objects movement unless the scene is perfectly still. In case of moving objects, the samples entering Equations (2.12 - 2.14) do not concern a specific scene point at subsequent instants as it should be, but different scene points at subsequent instants and expectedly cause distance measurement artifacts. The longer the integration time, the higher the likelihood of motion blur (but better the distance measurement precision). Integration time is another parameter to set in light of the specific application characteristics, needed for their imaging operation.

2.3 Matricial ToF cameras

Let us recall that the ToF sensors considered so far are single devices made by a single emitter and a co-positioned single receiver. Such an arrangement is only functional to single point distance measurements. The structure of actual ToF cameras is more complex than that of the ideal single ToF sensor cells considered so far, both because of the matrix nature of their ToF sensors and because of the optics needed for their imaging operation.

2.3.1 Matricial ToF sensors

A ToF camera sensor may be conceptually interpreted as a matricial organization of a multitude of single devices, each one made by an emitter and a co-positioned receiver as considered so far. In practice implementations based on a simple juxtaposition of a multitude of the previously considered single-point measurement devices are not feasible. Currently it is not possible to integrate $N_R \times N_C$ emitters and $N_R \times N_C$ receivers in a single chip, especially for high values of N_R and N_C as needed in imaging applications. However, it is not true that each receiver requires a specific co-positioned emitter, instead a single emitter may provide an irradiation that is reflected back by the scene and collected by a multitude of receivers close to each other. Once the receivers are separated from the emitters, the former can be implemented as CCD/CMOS lock-in pixels [32, 73] and integrated in a $N_R \times N_C$ matrix. The lock-in pixels matrix is commonly called *ToF camera sensor* (or simply *sensor*), and for example in the case of the MESA SR4000 it is made by 176×144 lock-in pixels.

Current matricial ToF sensor IR emitters are common LEDs that cannot be integrated. However they can be positioned in a configuration mimicking the presence of a single emitter co-positioned with the center of the receivers matrix, as shown in Figure 2.5 for the case of the MESA SR4000. Indeed the sum of all the IR signals emitted by this configuration can be considered as a spherical wave emitted by a single emitter, called *simulated emitter* (Figure 2.5), placed at the center of the emitters constellation. The fact that the actual emitters arrangement of Figure 2.5 is only an approximation of the non-feasible juxtaposition of single ToF sensor devices with emitter and receiver perfectly co-positioned introduces artifacts, among which a systematic distance measurement offset larger for the closer than for the further scene points. Figure 2.6 shows the actual emitters distribution of the MESA SR4000.

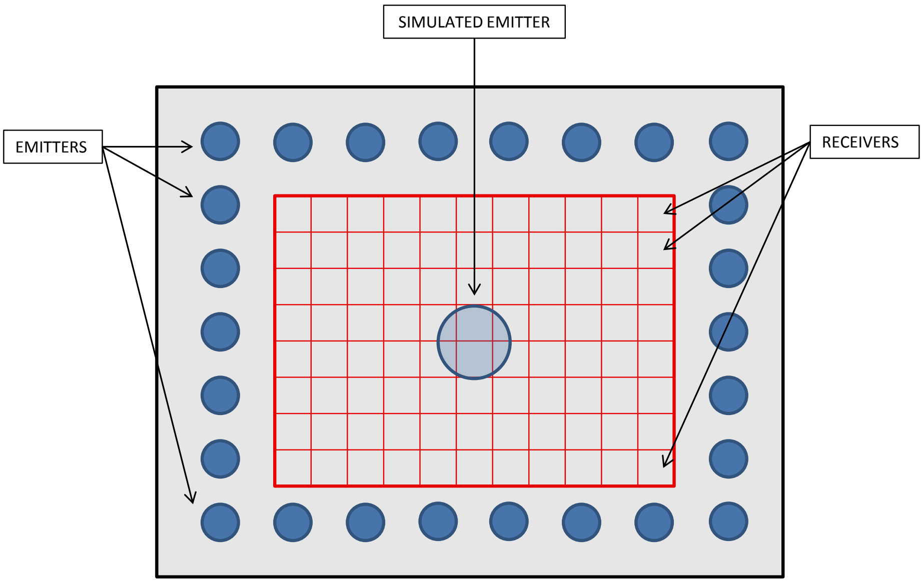


Figure 2.5: Scheme of a matricial ToF camera sensor. The CCD/CMOS matrix of lock-in pixels is in red. The emitters (blue) are distributed around the lock-in pixels matrix and mimic a simulated emitter co-positioned with the center of the lock-in pixel matrix (light blue).

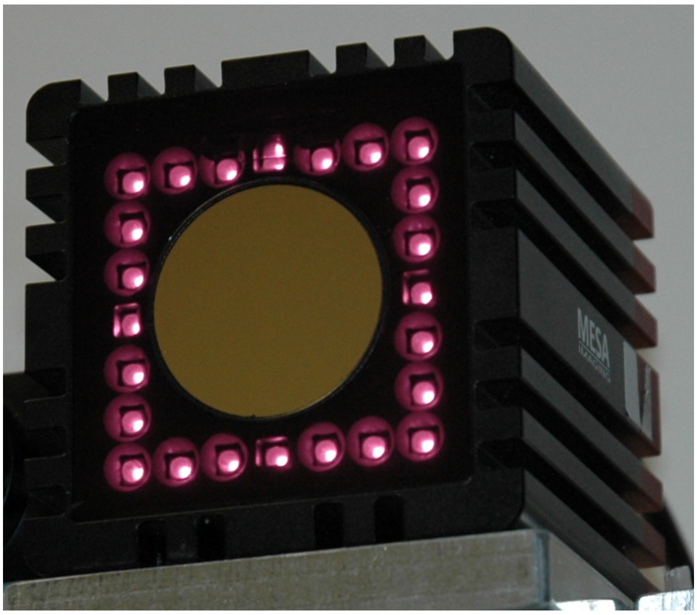


Figure 2.6: The emitters of the MESA SR4000 are the red LEDs.

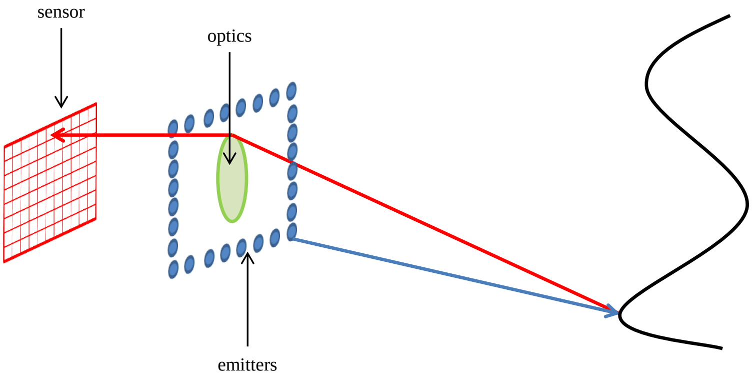


Figure 2.7: ToF camera structure and signaling: propagation towards the scene (blue arrow), reflection (from the black surface on the right), back-propagation (red arrow) towards the camera through the optics (green) and reception (red sensor).

2.3.2 ToF Camera imaging characteristics

ToF cameras can be modeled as pin-hole imaging systems since their structure, schematically shown in Figure 2.7, similarly to standard cameras, has two major components, namely the sensor made by a $N_R \times N_C$ matrix of lock-in pixels as explained in Section 2.3.1 and the optics.

ToF cameras, differently from standard cameras, have also a third important component, namely a set of IR emitters typically placed near the optics as shown in Figure 2.7. Figure 2.7 also shows that the IR signal sent by the emitters set travels toward the scene (blue arrow), it is reflected by the different scene portions, it travels back to the camera and through the optics (red arrow) it is finally received by the different lock-in pixels of the ToF sensor. The signaling process shown by Figure 2.7 is the basis of the relationship between the various scene portions and the respective sensor pixels.

All the pin-hole imaging system notation and concepts apply to ToF cameras. The notation presumes a pedix T in order to recall that it refers to a ToF camera. The camera coordinate system (CCS) of the ToF camera will be called the T-3D CCS. The position of a scene point with respect to the T-3D CCS will be denoted as P_T and its coordinates as $\mathbf{P}_T = [x_T, y_T, z_T]^T$. Coordinate z_T of P_T is called the *depth* of point P_T and the z_T-axis is called *depth axis*. The coordinates of a generic sensor pixel p_T of lattice Λ_T with the respect to the 2D-T reference system are represented by vector $\mathbf{p}_T = [u_T, v_T]^T$, with $u_T \in [0, ..., N_C]$ and $v_T \in [0, ..., N_R]$. Therefore the relationship

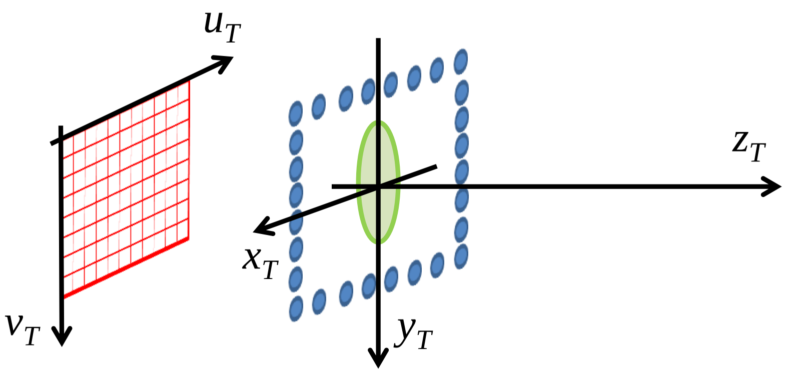


Figure 2.8: T-2D CCS (with axes $u_T - v_T$) and 3-3D CCS (with axes $x_T - y_T - z_T$).

between the 3D coordinates $\mathbf{P}_T = [x_T, y_T, z_T]^T$ of a scene point P_T and the 2D coordinates $\mathbf{p}_T = [u_T, v_T]^T$ of the pixel p_T receiving the IR radiation reflected by P_T is given by the perspective projection equation

$$z_T \begin{bmatrix} u_T \\ v_T \\ 1 \end{bmatrix} = K_T \begin{bmatrix} x_T \\ y_T \\ z_T \end{bmatrix} \tag{2.15}$$

where K_T is the ToF camera intrinsic parameters matrix [90].
Because of lens distortion, coordinates $\mathbf{p_T} = [u_T, v_T]^T$ of (2.15) are related to the coordinates $\hat{\mathbf{p}}_\mathbf{T} = [\hat{u}_T, \hat{v}_T]^T$ actually measured by the system by a relationship of type $\hat{\mathbf{p}}_\mathbf{T} = [\hat{u}_T, \hat{v}_T]^T = \Psi(\mathbf{p_T})$, where $\Psi(\cdot)$ is a distortion transformation. Such distortion can be modeled with standard parametrical approaches [35, 61]. The parameters of such model can be estimated with a camera calibration procedure, widely used also with ToF cameras.

As already explained, each sensor pixel p_T directly estimates the radial distance $\hat{r}_T$ from its corresponding scene point P_T. With minor and neglectable approximation due to the non-perfect localization between emitters, pixel p_T and T-3D CCS origin, the measured radial distance $\hat{r}_T$ can be expressed as

$$\hat{r}_T = \sqrt{\hat{x}_T^2 + \hat{y}_T^2 + \hat{z}_T^2} = \left|\left|[\hat{x}_T^2, \hat{y}_T^2, \hat{z}_T^2]^T\right|\right|_2 \tag{2.16}$$

From radial distance $\hat{r}_T$ measured at pixel p_T with distorted coordinates $\hat{\mathbf{p}}_T = [\hat{u}_T, \hat{v}_T]^T$ the 3D coordinates of $\mathbf{P}_T$ can be computed according to the following steps:

1. Given the lens distortion parameters, estimate the non-distorted 2D coordinates $\mathbf{p}_T = [u_T, v_T]^T = \Psi^{-1}(\hat{\mathbf{p}}_T)$, where $\Psi^{-1}(\cdot)$ is the inverse of $\Psi(\cdot)$;

2. The value $\hat{z}_T$ can be computed from 2.15 and 2.16 as

$$\hat{z}_T = \frac{\hat{r}_T}{\left\| K_T^{-1} \left[u_T, v_T, 1\right]^T \right\|_2} \tag{2.17}$$

   where K_T^{-1} is the inverse of K_T;

3. The values $\hat{x}_T$ and $\hat{y}_T$ can be computed by inverting (2.15), *i.e.*, as

$$\begin{bmatrix} \hat{x}_T \\ \hat{y}_T \\ \hat{z}_T \end{bmatrix} = K_T^{-1} \begin{bmatrix} u_T \\ v_T \\ 1 \end{bmatrix} \hat{z}_T \tag{2.18}$$

The operation of a ToF camera as imaging system can be summarized as follows. Each ToF camera sensor pixel, at each period of the modulation sinusoid, collects four samples s_R^0, s_R^1, s_R^2 and s_R^3 of the IR signal reflected by the scene. Every N periods of the modulation sinusoid, where N is a function of the integration time, each ToF sensor pixel estimates an amplitude value $\hat{A}$, an intensity value $\hat{B}$, a phase value $\widehat{\Delta\phi}$, a radial distance value $\hat{r}_T$ and the 3D coordinates $\hat{\mathbf{P}}_T = \left[\hat{x}_T, \hat{y}_T, \hat{z}_T\right]^T$ of the corresponding scene point.

Since amplitude $\hat{A}$, intensity $\hat{B}$ and depth $\hat{z}_T$ are estimated at each sensor pixel, ToF cameras handle them in matricial structures, and return them as 2D maps. Therefore a ToF camera, every N periods of the modulation sinusoid (which certainly correspond to several tens of times per second), provides the following types of data:

- An *amplitude map* $\hat{A}_T$, *i.e.*, a matrix obtained by juxtaposing the amplitudes estimated at all the ToF sensor pixels. It is defined on lattice Λ_T and its values, expressed in volt $[V]$, belong to the pixel non-saturation interval. Map $\hat{A}_T$ can be modeled as realization of a random field $\mathcal{A}_T$ defined on Λ_T, with values (expressed in volt $[V]$) in the pixel non-saturation interval.

- An *intensity map* $\hat{B}_T$, *i.e.*, a matrix obtained by juxtaposing the intensity values estimated at all the ToF sensor pixels. It is defined on lattice Λ_T and its values, expressed in volt $[V]$, belong to the pixel non-saturation interval. Map $\hat{B}_T$ can be modeled as realization of a random field $\mathcal{B}_T$ defined on Λ_T, with values (expressed in volt $[V]$) in the pixel non-saturation interval.

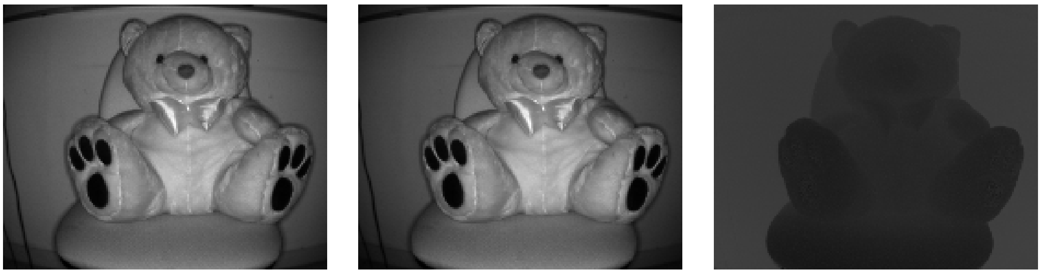


Figure 2.9: Example of $\hat{A}_T$, $\hat{B}_T$ and $\hat{Z}_T$ (in this order from left to right in the figure).

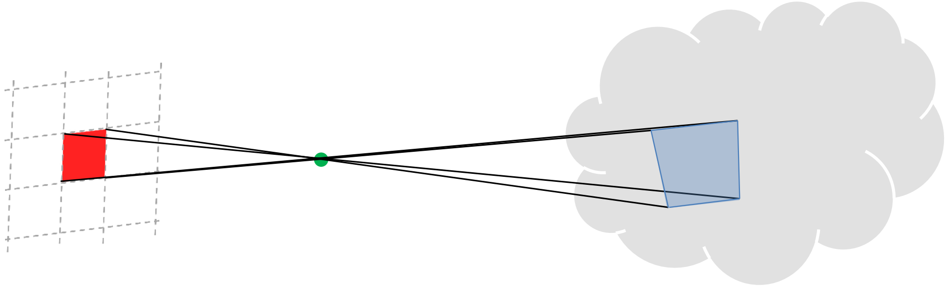


Figure 2.10: Finite size scene area (blue) associated to a ToF sensor pixel (red).

- A *depth map* $\hat{Z}_T$, i.e, a matrix obtained by juxtaposing the depth values estimated at all the ToF sensor pixels. It is defined on lattice Λ_T and its values, expressed in $[mm]$, belong to interval $\left[0, r_{MAX} = \frac{c}{2f_{mod}}\right)$. Map $\hat{Z}_T$ can be considered as realization of a random field $\mathcal{Z}_T$ defined on Λ_T, with values (expressed in $[mm]$) in $[0, r_{MAX})$.

By normalizing amplitude, intensity and depth values into the interval $[0, 1]$ the three maps $\hat{A}_T$, $\hat{B}_T$ and $\hat{Z}_T$ can be represented as images as shown in Figure 2.9 for a sample scene. For the scene of Figure 2.9 images $\hat{A}_T$ and $\hat{B}_T$ are very similar because the scene illumination is rather constant.

2.3.3 Practical imaging issues

As expected the actual imaging behavior of ToF cameras is more complex than that of a simple pin-hole system and some practical issues must be taken into account. First of all, it is not true that a sensor pixel is associated to a single scene point, but it is associated to a finite scene area, as shown in Figure 2.10. For this reason, each pixel receives the radiation reflected from all the points of the corresponding scene area. If

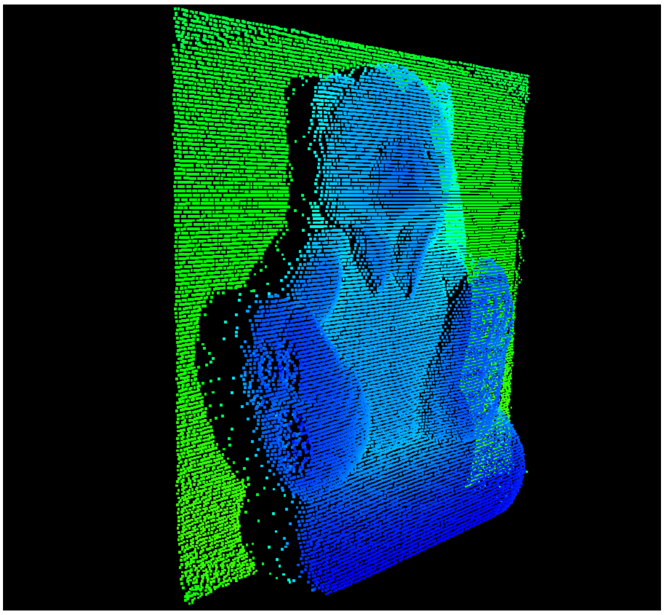


Figure 2.11: An example of flying pixels at the depth edge between object and wall.

the scene area is a flat region with somehow constant reflectivity, the approximation that there is a single scene point associated to the specific pixel does not introduce any artifact. However, if the area crosses a reflectivity discontinuity, the values of $\hat{A}_T(p_T)$ and $\hat{B}_T(p_T)$ estimated by the correspondent pixel p_T average somehow its different reflectivity values. A worse effect occurs if the area associated to p_T crosses a depth discontinuity. In this case assume that a portion of the area is at closer depth, called z_{near}, and another portion at further depth, called z_{far}. The resulting depth estimate $\hat{Z}_T(p_T)$ is a convex combination of z_{near} and z_{far}, where the combination coefficients depend on the percentage of area at z_{near} and at z_{far} respectively reflected on p_T. The pixels associated to such depth estimates are commonly called *flying pixels*. The presence of flying pixels leads to severe depth estimation artifacts, as shown by the example of Figure 2.11.

Multi-path propagation is a major interference in ToF camera imaging. As shown in Figure 2.12, an optical ray (red) incident to a non-specular surface is reflected in multiple directions (green and blue), a phenomenon commonly called *scattering*. The ideal propagation scenario with co-positioned emitters and receivers considers only the presence of the green ray of Figure 2.12, *i.e.*, the ray back reflected in the direction of the incident ray and disregards the presence of the other (blue) rays. In practical

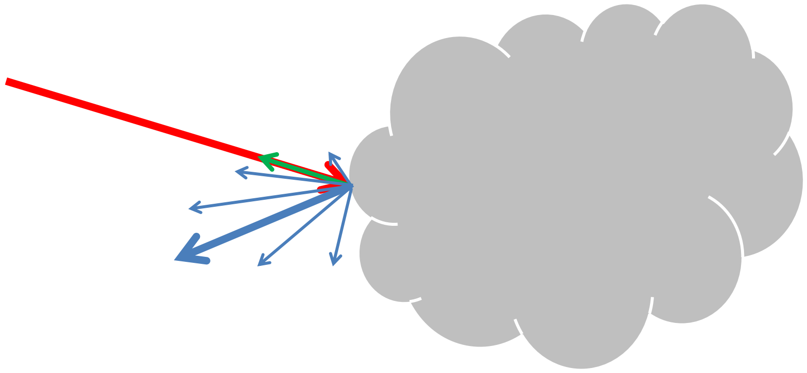


Figure 2.12: Scattering effect.

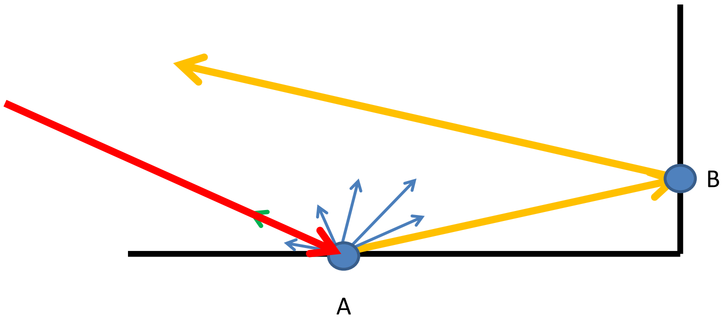


Figure 2.13: Multi-path phenomenon: the incident ray (red) is reflected in multiple directions (blue and orange rays) by the surface at point A. The orange ray reaches then B and travels back to the ToF sensor.

situations, however, the presence of the other rays may not always be neglectable. In particular, the ray specular to the incident ray direction with respect to the surface normal at the incident point (thick blue ray) is generally the reflected ray with greatest radiometric power. All the reflected (blue) rays may first hit others scene points and then travel back to the ToF sensor, affecting therefore the distance measurements of other scene points. For instance, as shown in Figure 2.13, an emitted ray (red) may be firstly reflected by a point surface (A) with a scattering effect. One of the scattered rays (orange) may then be reflected by another scene point (B) and travel back to the ToF sensor. The distance measured by the sensor pixel relative to B is therefore a combination of two paths, namely path to ToF camera - B - ToF camera and path ToF camera-A-B-ToF camera. The coefficients of such a combination depend on the optical amplitude of the respective rays. Since the radial distance of a scene point P from the

ToF camera is computed from the time-length of the shortest path between P and the ToF camera, the multi-path effect leads to over-estimate the scene points distances.

Multi-path is one of the major error sources of ToF cameras distance measurements. Since multi-path is scene dependent it is very hard to model. Currently there is no method for its compensation, but there are practical provisions that might alleviate the multi-path effects, as explained in [65].

3

Stereo vision systems

A stereo system exploits the images from a pair of standard video-cameras in order to provide an estimate of the depth distribution of the scene framed by the two cameras. All stereo vision systems are based on the triangulation principle: given two cameras pointing towards an object, the difference between the positions of the object in the two acquired images is inversely proportional to the distance of the object from the cameras (as later formalized in this chapter). Two examples of commercial stereo vision systems are the one by Point Gray [17] and the one by TZYX [20], recently sold to Intel [7].

The *hardware component* of the stereo system is made by a pair of standard video-cameras and optionally by a synchronization circuit rather useful in case of dynamic scenes. Depth information computed by stereo systems is relative to the point of view of one of the two cameras, usually called the reference camera, while the other one is usually called target camera. In this thesis the reference camera will be the left one (denoted by L) and the target the right one (denoted by R). The acquired images are called either reference or target images depending on the camera acquiring them (or also left and right images). For each camera a 2D CCS is associated in order to describe the coordinates of pixels in the acquired images and a 3D CCS is associated in order to describe the positions of scene points with respect to the camera itself. For the left camera such CCS are respectively called L-2D CCS and R-2D CCS. For the right camera they are called R-2D CCS and R-3D CCS. The CCS of the left camera are usually adopted as stereo coordinates system. A schematic representation of such CCS is shown in Fig. 3.1.

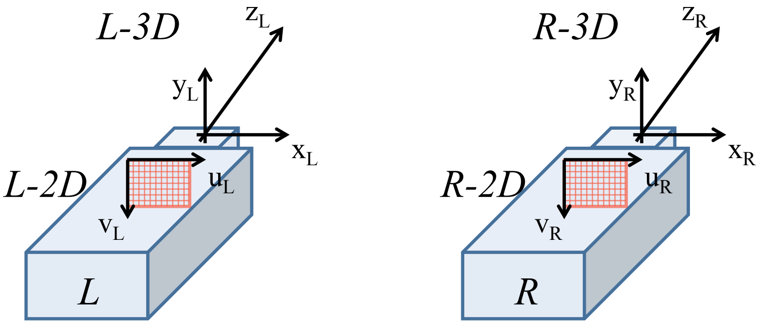


Figure 3.1: Schematic representation of the 2D CCS and of the 3D CCS associated to the left and the right camera.

The estimation of the relationships between these four CCSs is obtained via stereo calibration algorithms [30, 90]. The output of such calibration procedure is the estimates of all the parameters describing the projection properties of the two cameras, *i.e.*, the intrinsic camera parameters and the roto-translation between the two 3D CCS, *i.e.*, the extrinsic parameters. The intrinsic parameters are generally represented in a matricidal form by the matrix of intrinsic parameters (K_L for L and K_R for R). The extrinsic parameters are represented as a rotation matrix R_S and a translation vector T_S (where S stays for stereo).

Given a calibrated stereo system, it is it is customary to apply a rectification procedure to the images acquired by the two cameras in order to simplify the task of stereo vision algorithms. Rectification takes as input the images acquired by L and R and performs the following operations:

1. Correction of the projective distortion introduced by the camera lenses

2. Compensation of the focal length differences between L and R

3. Compensation of the differences in the other intrinsic parameters of the L and R cameras.

4. Compensation of the relative rotation between the two cameras in order to obtain images as if they were acquired by cameras with parallel optical axes orthogonal to the line through the optical center of L and R.

For details on rectification, the reader is referred to [30, 55, 90].

An image acquired by L, after rectification is called rectified reference image (or rectified left image), and denoted by I_L. An image acquired by R, after the rectification process is called rectified target image (or rectified right image), and denoted by I_R. The two images I_L and I_R are associated each one to a standard 2D reference system, with horizontal axis u pointing rightward and vertical axis v pointing downward.

It is worth pointing that for a rectified stereo system, no rotation is assumed between the 3D CCSs associated with L and R as well as no translation along the y and z directions. Scene point P_S with coordinates $\mathbf{P}_S = [x_S, y_S, z_S]^T$ expressed with respect to the L-3D CCS, if visible from both cameras, is projected to point p_L with coordinates $\mathbf{p}_L = [u_L, v_L]^T$ on I_L expressed with respect to the L-2D CCS and to point p_R with coordinates $\mathbf{p}_R = [u_R, v_R]^T = [u_L - d, v_L]^T$ on I_R expressed with respect to the R-2D CCS. It can be shown that the difference d between the coordinates of the two 2D points, called disparity, and the depth value z of P is

$$d = \frac{bf}{z} \tag{3.1}$$

where b is the baseline, *i.e.*, the distance between the nodal points of L and R, and f is the focal length (equal for both rectified cameras). Points p_L and p_R, called conjugates because of rectification, share the same vertical coordinate v. One can associate a disparity value to each pixel p_L and obtain an image of disparity values, denoted as D_S and called disparity image or disparity map. From (3.1) two observations are in order: high values of d correspond to points close to the cameras, *i.e.*, to points with low z value; since d is generally quantized and there is an inverse relationship between z and d, the accuracy of the stereo vision systems does not decrease linearly, but quadratically with respect to z according to

$$\Delta z = \frac{z^2}{fb}\Delta d \tag{3.2}$$

where Δd is the disparity quantization step and Δz the depth quantization step. Because of rectification only non-negative values of d are valid and $d = 0$ corresponds to points with depth value $z = \infty$. It is customary to limit the range the values d may take from geometrical considerations. If the minimum and the maximum depth values (respectively z_{MIN} and z_{MAX}) of the scene are known, the disparity excursion, can be confined to $d \in [d_{MIN}, d_{MAX}]$, with $d_{MIN} = bf/z_{MAX}$ and $d_{MAX} = bf/z_{MIN}$.

3.1 Stereo vision algorithms

It has been shown in the previous section that for a rectified stereo system, the value of depth distribution z of a scene point P with coordinates $\mathbf{P} = [x, y, z]^T$ visible from both cameras can be obtained by (3.1) from the estimation of disparity distribution d between all the pairs of conjugate points $p_L \in I_L$ with coordinates $\mathbf{p}_L = [u, v]^T$ and $p_R \in I_R$ with coordinates $\mathbf{p}_R = [u - d, v]^T$. Hence the information about the depth distribution of a scene is coded by the *disparity image* D_S, which is a typical intermediate output of stereo algorithms. The computation of the depth distribution a scene is called *computational stereopsis* or *triangulation* [90] and encompasses two steps, the first is a point matching procedure corresponding to a linear search meant to detect conjugate points along each horizontal line of I_L, row by row and the second is the computation of the depth distribution z from the disparity image D_S by (3.1). Point matching is a rather critical step since wrong matches inevitably lead to wrong scene depth estimates. Stereo matching can be performed in many ways, essentially trading speed against robustness, and it is a distinctive element differentiating the various stereo algorithms. A wide class of stereo algorithms, called *local methods*, exploits local similarity in order to detect, given $p_L \in I_L$, the point p_R on the corresponding line of I_R with neighborhood most similar to that of p_L (of course similarity can be defined in many ways). Other algorithms, called *global methods*, adopt a global model of the scene, by implicitly or explicitly imposing constraints on the overall scene depth configuration. *Semi-global methods* use scene models imposing constraints only on parts of the scene depth. The following subsections review three examples of these methods currently in great consideration and usage, namely the most classical local algorithms, *i.e.*, Fixed Window (FW); a widely adopted global algorithm, *i.e.*, Loopy Belief Propagation (LBP); and Semi Global Matching (SGM), a state of the art semi-global algorithm.

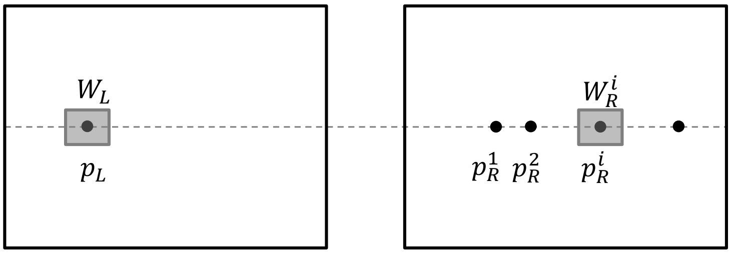


Figure 3.2: Fixed Window (FW) stereo algorithm.

3.2 Local stereo algorithms

The Fixed Window (FW) stereo algorithm is a classical local algorithm widely used in practical implementations for its simplicity. For each pixel $p_L \in I_L$ with coordinates $\mathbf{p}_L = [u, v]^T$ its conjugate $p_R \in I_R$ with coordinates $\mathbf{p}_R = [u - d^*, v]^T$ (and equivalently its disparity value d^*) is computed as follows:

- A squared (or rectangular) window W_L is centered around p_L and other windows of the same size W_R^i are centered around each candidate conjugate point $p_R^i(u - i, v), i = 1, ..., d_{MAX} - d_{MIN}$ as shown in Fig. 3.2.

- Cost c_i of matching p_L against each one of the candidate conjugate points p_R^i is computed by comparing I_L on W_L and I_R on each W_R^i. An example of such a costs and type of comparisons is the Sum of Absolute Differences (SAD), i.e,

$$c_i = \frac{1}{|W_L|} \sum_{p \in W_L, q \in W_R^i} |I_L(p) - I_R(q)| \tag{3.3}$$

  where $|W_L|$ is the number of pixels in W_L and p and q are characterized by the same position in the relative windows. Clearly many other different measures could be used in this task, *e.g.*, the correlation, the sum of squared differences or the census transform [86].

- Pixel p_R^i corresponding to the minimum matching cost c_i is selected as conjugate of p_L as well as the estimated disparity $d^* = d_i$.

Such a local method considers a single pixel of I_L at the time, it adopts a Winner-Takes-All (WTA) strategy for disparity optimization, and it does not explicitly impose any model on the depth distributions. Like most local approaches, cost aggregation within fronto-parallel windows implicitly assumes the same disparity for all the points within the window. This is clearly not true if the window includes a scene depth discontinuity. Indeed FW is well known not to perform well across depth discontinuities. Moreover, as most local algorithms, FW performs poorly in texture-less regions. Nevertheless, since incremental calculation schemes, *e.g.* [38, 78], can make FW very fast, it is widely used in practical applications despite its notable limitations. The larger the window size the better the robustness against image noise and low texture situations, at the expense of the precision in presence of discontinuities.

Evolutions of FW focus on the shape of the coupling window [68], on the usage of multiple coupling windows for a single pair of candidate conjugate points [56], on weighting the contribution of the different pixels within a window according to suitable weights, given for instance by a bilateral filter [66] or derived from segmented versions of I_L and I_R [92]. These modifications of the classical fixed window strategy improve its performance, especially in presence of depth discontinuities, but significantly increase computation/execution time. An interesting variant of FW applies the SAD strategy to color images I_R and I_L (assumed available) by separately treating their color channels.

3.3 Global stereo algorithms

While local stereo algorithms estimate the disparity image D_S almost independently for each pixel by a WTA strategy applied to costs computed on local portions of the reference and target images, global stereo vision algorithms compute the whole disparity image D_S at once by imposing a smoothness model on scene depth distribution. Such global algorithms generally adopt a Bayesian framework and model the disparity image as a *Markov Random Field* (MRF) in order to include within a unique framework cues coming from local comparisons between reference and target image and smoothness constraints. Global stereo vision algorithms typically estimate the disparity image by minimizing a cost function made by two terms:

$$\hat{I}_D = \arg\min_D [C_{data}(I_L, I_R, D_S) + C_{smooth}(D_S)] \tag{3.4}$$

The quantity $C_{data}(I_L, I_R, D_S)$ is the so-called *data term* , representing the cost of a local matches (similar to the one of local algorithms). The sum of such costs over all the reference image points defines the cost of a disparity image D_S. This term encodes the same type of information contained in the cost term of local stereo algorithms in Equation (3.3).

The quantity $C_{smooth}(D_S)$, called *smoothness term*, defines the level of smoothness of disparity image D_S, by explicitly or implicitly accounting for discontinuities. The term $C_{smooth}(D_S)$ takes into account that scenes generally have quite flat disparity distributions except in presence of depth discontinuities, by penalizing disparity images that do not respect this type of behavior. With a MRF model of the disparity image, $C_{smooth}(D_S)$ can be computed as sum of local terms accounting for the smoothness of neighboring pixels. Equation 3.4 can be obtained as the final expression of a *Maximum-A-Posteriori* (MAP) formulation of the stereo problem. Other terms can be added to Equation (3.4), in order to explicitly model occlusions and other a-priori knowledge on the scene depth distribution.

Minimization (3.4) is not trivial, because of the great number of variables involved, *i.e.*, $n_{row} \times n_{col}$ disparity values of D_S, which can assume $d_{MAX} - d_{MIN} + 1$ possible values within range $[d_{MIN}, d_{MAX}]$. Therefore there are $(n_{rows} \times n_{cols})^{d_{MAX} - d_{MIN} + 1}$ possible configurations of D_S. Since images acquired by current cameras can easily have millions of pixels within the range of hundreds of values, it is easy to understand how a greedy search for the minimum over all the possible configurations of D_S is not feasible. A classical solution to this is Loopy Belief Propagation (LBP), which searches for the minimum cost solution of (3.4) in a probabilistic sense. The disparity image is considered as a MRF made by the juxtaposition of random variables (one for each pixel in D_S). Instead of optimizing the global probability density function defined on the whole random field, LBP marginalizes it, obtaining a probability density function for the disparity distribution of each point of D_S. The final optimization is performed by independently maximizing the marginalized probability density function at each point of D_S. The application of LBP to stereo vision has been proposed in [89]. An extensive description of LBP can be found in [24, 76]. An interesting perspective for the algorithms used for solving huge problems such as minimization (3.4) can be found in [91].

Global stereo vision algorithms are typically way more computational expensive than local algorithms. However, by explicitly modeling the smoothness constraints (and by possibly including other constraints), they are able to cope with depth discontinuities and are more robust in texture-less regions.

3.4 Semi-global stereo algorithms

Another very interesting class of stereo algorithms is constituted by the semi-global stereo approaches, which similarly to global methods adopt a global disparity model, but differently than global methods in order to reduce the computational burden do not compute it on the whole disparity image. More precisely the minimization of the cost function is computed on a reduced model for each point of D_S, differently than global approaches which estimate a whole disparity image D_S at once. For instance, the simplest semi-global methods, such as Dynamic Programming or Scanline Optimization [37] work in a 1D domain and optimize each horizontal image row by itself. The so-called Semi Global Matching (SGM) algorithm [64] is a more refined semi-global stereo algorithm. It explicitly models the 3D structure of the scene by means of a point-wise matching cost and a smoothness term. Several 1D energy functions computed along different paths are independently and efficiently minimized, and their costs are summed up. For each point, the disparity corresponding to the minimum aggregated cost is selected. In [64] the authors propose to use 8 or 16 different independent paths. The SGM approach works well near depth discontinuities, however, due to its (multiple) 1D disparity optimization strategy, produces less accurate results than more complex 2D disparity optimization approaches. Despite its memory footprint, this method is very fast and potentially capable to deal with poorly textured regions.

4

Comparison of ToF cameras and stereo systems in terms of metrological quantities

In this chapter a set of tools for measuring the quality of depth information acquired by a ToF camera or by a stereo system is introduced. Such tools are also applied in order to describe ToF cameras and stereo systems and to assess their complementarity. Let us first briefly recall the concepts of accuracy, precision and measurement resolution. For a detailed presentation, the reader is referred to [5, 31]. Consider a measurement system S measuring a physical quantity Q. Assume the actual value of Q to be q^*. System S performs a series of n independent measurements of Q, all under the same experimental conditions. The values measured by S at each step are: $q_1, q_2, , q_N$.

Definition 1. *The accuracy A of a measurement system S is the degree of closeness of measurements q_n to the actual value q^* of the quantity Q. It can be computed as the difference between the average on a set of measures of the same quantity and the actual value, i.e. $A = |\bar{q} - q^*|$, where $\bar{q} = \frac{1}{N}\sum_{n=1}^{N} q_n$.*

In the specific case of acquisition of depth maps, *i.e.*, of depth information $z(p_{i,j})$ organized as an $I \times J$ matrix, as produced by ToF cameras and stereo vision systems, assume there are $z^n(p_{i,j})$, $n = 1, 2, , N$ depth map measurements of the scene Q available. In this case the accuracy of the measurement system is defined as

$$A = \frac{1}{I \times J} \sum_{i=1}^{I} \sum_{j=1}^{J} |\bar{z}(p_{i,j}) - z^*(p_{i,j})| \tag{4.1}$$

where $\bar{z}(p_{i,j}) = \frac{1}{N}\sum_{n=1}^{N} z(p_{i,j})$ and $z^*(p_{i,j})$ is the ground truth depth map.

Definition 2. *The precision (or repeatability) P of a measurement system S is the degree to which repeated measurements under unchanged conditions show the same result. A common convention is to calculate the precision P of the system S in the measure of Q as the standard deviation of the measurement distribution σ_q of the measurements $q_1, q_2, , q_N$, i.e., $P = \sqrt{(\frac{1}{N}\sum_{n=1}^{N}(q_n - \bar{q})^2}$, where $q = \frac{1}{N}\sum_{n=1}^{N} q_n$.*

The precision of a depth acquisition system can be computed by performing several depth measurements $z^n(p_{i,j}), n = 1, 2, , N$ and computing the standard deviation averaged over the whole depth map

$$P = \frac{1}{I \times J}\sum_{i=1}^{I}\sum_{j=1}^{J}\sqrt{\frac{1}{N}\sum_{n=1}^{N}(z^n(p_{i,j}) - \bar{z}(p_{i,j}))^2} \tag{4.2}$$

where $\bar{z}(p_{i,j})$ is defined as above.

Definition 3. *The measurement resolution R of a measurement system S is the smallest change δ_q in the underlying physical quantity with actual value q^* that produces a response in the measurement system.*

The resolution of a depth acquisition system can be further specified into *lateral resolution*, *i.e.*, the amount of pixels in the sensor $(I \times J)$ and *depth resolution*, *i.e.*, the minimum amount of difference in depth that the system is able to recognize.

4.1 Accuracy, precision and resolution of ToF cameras and stereo systems

In this sections the previously introduced quantities are considered in the special case of ToF cameras and stereo systems. From the following analysis it is clear that ToF cameras and stereo systems are rather different with respect to accuracy, precision and measurement resolutions

4.1.1 Accuracy

With respect to accuracy, it is well known that ToF cameras depth measurements are characterized by a systematic offset caused by the harmonic distortion of the illuminators and camera pixels circuitry which generally varies with the distance and can be up to some hundreds of millimeters (*e.g.*, $400[mm]$, as reported in [67] and shown in Figure 2.4). In order to account for this artifact, one should provide an accuracy value for each distance value in the range of the measurable distances (*e.g.*, in $500 - 5000[mm]$). For system characterization purposes it is customary to synthesize the accuracy by a single value obtained by averaging the accuracy of the instrument over the range of measurable distances. The ToF depth measurement offset due to harmonic distortion is of systematic nature and it can be reduced by a Look-Up-Table (LUT) correction independently applied to each pixel. However, since the measurement error also depends on the scene geometry and reflectance distribution, the LUT correction does not completely cancel the measurement error. The LUT-improved accuracy of a ToF camera is therefore limited. For example, according to the producer, the MESA SR4000 [8] is characterized by an accuracy of about $10[mm]$.

The accuracy of stereo vision systems depends on the texture and geometry characteristics of the acquired scene. The great variability of possible geometry and textures leads to non-systematic measurement errors which cannot benefit from simple strategy such as LUT-compensation. In order to better understand the origin of stereo systems error, let us consider the case of the FW stereo algorithm, which for each point in the reference image identifies a conjugate point in a segment on the epipolar line in the target image. Each couple of candidate conjugate points is characterized by a matching likelihood, quantified by a cost function (*e.g.*, TAD). The more the two images are similar near to candidate conjugate points, the lower is their cost function and the more likely is the matching. The best case for stereo vision systems is when the scene characteristics are such that the local similarity between the L and the R images is high only in correspondence of the actual conjugate points pair (and low for the other candidate points pairs). In such a case, the cost function has a minimum in correspondence of the conjugate points pair actually estimated by the WTA algorithm. This lucky situation requires that the reference and the target image satisfy the following two conditions:

- Reference and target image should exhibit an adequate amount of color information (texture) near the actual conjugate points pair (aperture problem)
- No other region of the target image along the epipolar line should be similar to the one corresponding to the actual target conjugate point (repetitive texture pattern)

In case of insufficient texture or of multiple candidate conjugate points locally similar to the local reference image, there might be a disparity estimation mismatch with a consequent depth estimation error. Scene illumination greatly influences the possibility of this type of mismatches. Such depth measurement error does not grow regularly with the image noise, but tends to sudden bursts when the scene characteristics make the system unable to find the correct cost function minimum. The accuracy of the depth measurements produced by a stereo system is very hard to characterize by a single parameter since it strongly depends on the scene characteristic and on the considered stereo vision algorithm. All one can do is to define the accuracy of a stereo vision system for a specific scene or specific reference objects under specific illumination conditions (from different acquisitions of the same scene under the same conditions, and by computing the difference between the averaged estimated depth value at each pixel with respect to its actual depth value) as shown in Section 4.2. In general local stereo algorithms, totally dependent from the scene color distribution with respect to accuracy they perform poorer than global and semi-global techniques which are less dependent on scene characteristics, because of the assumed smoothness model. At the same time it is clear that in case the actual scene does not match the assumed model, the assumptions behind global and semi-global methods turn against performance accuracy.

4.1.2 Precision

According to Chapter 2 the noise of ToF depth measurements can be approximated to be Gaussian [67]. The depth measurement accuracy of ToF cameras relates directly to the mean of this Gaussian process, while the depth measurement precision is defined as its standard deviation that can be computed according to Equation (2.11). The standard deviation of the measurements increases as the distance from the object or the background illumination increase or the object reflectance decreases. For instance in the case of high reflectivity targets and low IR background illumination, the precision of the MESA SR4000 [8], according to the producer, is less than $20[mm]$.

For the analysis of the precision of stereo systems let us consider the simple FW stereo algorithm. For simplicity denote by $Z^l = Z^l(p_{i,j}), i = 1, 2, ..., I, j = 1, 2, ..., J$ the l-th depth map measurement. Assume, as shown in Figure 4.1 that the scene is acquired N times under same conditions giving the N images $I_L^1, I_L^2, ..., I_L^N$ from L, the N images

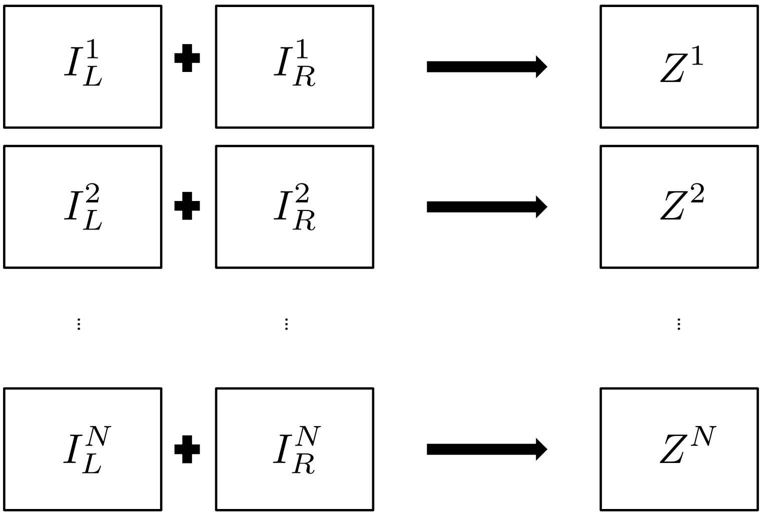


Figure 4.1: Schematic representation of acquired stereo images and relative depth maps.

$I_R^1, I_R^2, ..., I_R^T$ from R from which the N corresponding depth maps $Z^1, Z^2, ..., Z^N$ are computed by the FW stereo vision algorithm.

The N depth maps are usually similar, but not identical due to the noise affecting images $I_L^1, I_L^2, ..., I_L^N$ and $I_R^1, I_R^2, ..., I_R^N$. Hence for a given point p_L the matching cost with respect to each candidate conjugate points varies for each acquisition. Noise fluctuations changing the conjugates pair that minimizes the matching cost change also the estimated depth map. The noise amount needed for changes of this nature clearly depends on the amount of texture in the scene. Low textured scenes are highly affected by image acquisition noise, while high textured scenes are less affected by it. The precision of FW stereo algorithm is directly related to scene reflectance characteristics and illumination conditions. Other stereo algorithms, such as SGM and BP are less noise-prone than FW, because the imposed scene model generally capable to mitigate the noise influence. The precision of a stereo vision system, with respect to a specific scene or reference object can be obtained from N acquisitions (as shown in Figure 4.1) by computing the standard deviation of the measurements for each depth map point according to (4.2).

4.1.3 Resolution

The measurement resolution of a matricial depth acquisition system, such as ToF cameras and stereo systems is characterized by spatial and depth resolution. The spatial resolution (or lateral resolution) for a fixed field-of-view (uniquely identified by the optics) is determined by the number image pixels and it represents the measurements resolution in the $x - y$ scene coordinates. The depth resolution, or resolution in the scene z coordinates, is the smallest scene variation δ_z capable to produce a depth response. The spatial resolution of ToF cameras, *i.e.*, the number of pixels in the sensor matrix, is currently considered one of their limitations, and it is one of the targets of ToF technology advancement. For instance, in the case of the MESA SR4000, the sensor matrix has 176×144 pixels. The analysis of a ToF camera depth resolution can be experimentally made as follows. Consider a set of N measurements of the ToF camera T positioned at a known distance z from a reference object, typically a plane of metrologically known characteristics. The minimum depth difference $\delta_Z(z)$ that produces a noticeable difference in the average of the depth measurements of two depth measures is the depth resolution of the camera T. Various factors may influence $\delta_Z(z)$, *i.e.*, the sensitivity of the ToF cameras pixels, the precision of the sensor hardware and the final quantization grain of the depth measurements. Such a quantization grain is usually very fine. For example, the MESA SR4000 samples a depth interval of $5000[mm]$ with 2^{14} values, *i.e.*, with a quantization step of $0.3[mm]$. The other elements conditioning depth resolution cannot be treated analytically, and depth resolution must be estimated. As a practical example, the ToF resolution, for instance at $z = 1000[mm]$, can be measured by taking a planar object and moving it from z to $z + \delta_z$ for smaller and smaller values of δ_z and by taking N measurement for each value of δ_z (*e.g.* with $N = 10^5$). If, for instance, at $\delta_z < 1[mm]$ the average of the ToF measurements at z and $z + \delta_z$ coincides and at $\delta_z^I = 1[mm]$ they do not coincide, it is possible to state that $\delta_z^I = 1[mm]$ is the resolution. In the case of stereo systems the analysis of the measurements resolution can be done analytically. The spatial resolution of a stereo vision system is just given by the number of pixels of the left camera image sensor matrix. Since such matrices have a great number of pixels (*e.g.*, 1032×778) stereo systems are considered high spatial resolution systems. This is certainly true, but it is also important to remind that stereo systems cannot estimate the depth value of all the points in their images and especially in presence of depth discontinuities they are not very precise. Furthermore it is possible to compute a disparity value only for samples visible by both cameras, *e.g.*, usually the disparity cannot be estimated for the first columns on the side of the image or for points occluded with respect to one of the two cameras. Concerning the depth resolution of stereo vision systems it is important to recall from Equation (3.1) that the relationship between disparity and depth is not linear. Since the disparity is linearly sampled (the disparity for each pixel is an integer in the interval $[d_{MIN}; d_{MAX}]$), the relative depth values are non-linearly sampled. Furthermore the quadratic dependence from depth values z of the depth increments Δz given by Equation (3.2) has important consequences on depth resolution. Suppose that a point at depth z^* is acquired by a stereo system characterized by a focal f and a baseline b. The actual disparity value of that point is $d^* = \frac{bf}{z^*}$. The estimate $\hat{d}$ of d^* assumes only an integer value in $[d_{MIN}, d_{MAX}]$ which will be either $\lfloor \hat{d} \rfloor$ if $\hat{d} - \lfloor \hat{d} \rfloor 0.5$, or otherwise $\lceil \hat{d} \rceil$. Consequently the estimate $\hat{z}$ of z might assume either value $\hat{z} = \frac{bf}{\lfloor \hat{d} \rfloor}$ or $\hat{z} = \frac{bf}{\lceil \hat{d} \rceil}$ and the minimum depth increment that the system can measure for a point at distance z^* is $\Delta z = \frac{bf}{\lfloor d \rceil} - \frac{bf}{\lceil d \rceil}$. From Equation (3.2) $\Delta z = \frac{z^{*2}}{fb} \Delta d$, where in this case $\Delta d = \lceil d \rceil - \lfloor d \rfloor = 1$. In other words the depth resolution decreases quadratically with the depth of the measured objects. Depth resolution can be improved by sub-pixel stereo matching, but the benefits are limited by interpolation artifacts. Sub-pixel techniques allow to reduce the value of Δd in Equation (3.2) (*e.g.* $\Delta d \approx 0.1$), but cannot change the quadratic dependence of Δz with respect to depth z. Therefore ToF cameras usually have a better depth resolution Δz than stereo systems for distant objects and worse resolution than stereo systems for close objects. Another important element to take into account is the computation time of the different scene depth estimation systems. While ToF cameras operation is very simple and can be efficiently implemented in hardware, stereo algorithms, especially the global ones are computational complex. Rates of tents of depth estimates per second (*e.g.*, 50 times per second) are typical of ToF cameras, while rates of few depth estimates per second are typical of software implementations of current stereo algorithms. Needless to say, the speed of stereo vision algorithms can be greatly improved by hardware implementations.

4.2 Experimental comparison

In order to clarify the previous discussion some experimental comparisons of the performances of ToF cameras and stereo vision systems on a sample dataset are presented next. The reference scene showed in Figure 4.2 has been acquired by both the ToF camera and the stereo acquisition system of the setup shown in Figure 5.1. The scene depth map has then been estimated by three different stereo vision algorithms, namely, FW, SGM and LBP. The goal of this experiment is to give an example of how comparisons of this kind can be made in practice.

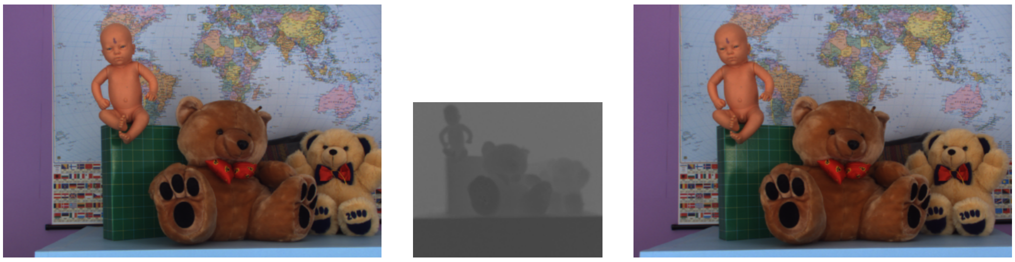


Figure 4.2: Undistorted data acquired by the trinocular acquisition system of Figure 5.1 made by a stereo acquisition system and a ToF camera. Starting from the left, the color image I_L acquired by L, the depth map Z_T acquired by T and the color image I_R acquired by R are shown.

The implementations of the considered stereo vision algorithms can be found in the OpenCV library [13]. In particular, FW and SGM implementations are classical CPU stereo vision algorithms, while the considered LBP implementation exploits also the GPU. A matching window of size 21×21 has been adopted for FW and SGM stereo vision algorithms, while a small 1×1 window for LBP. The scene was acquired $N = 10$ times by both the stereo system and the ToF camera. The three considered stereo vision algorithms have been applied to each stereo acquisition. In order to have a ground truth depth measurement the scene was also acquired by an active space-time stereo vision system [49, 101], with an accuracy of about $3 - 4[mm]$, way superior to that of both the stereo and the ToF camera. Figure 4.3 shows three examples of estimated depth maps (one for each stereo algorithm) and the ground-truth depth-map computed by the space-time stereo.

Note that a depth measurement is not available for the pixels associated to zero depth (black pixels) due to matching failure and occlusions in the case of the passive

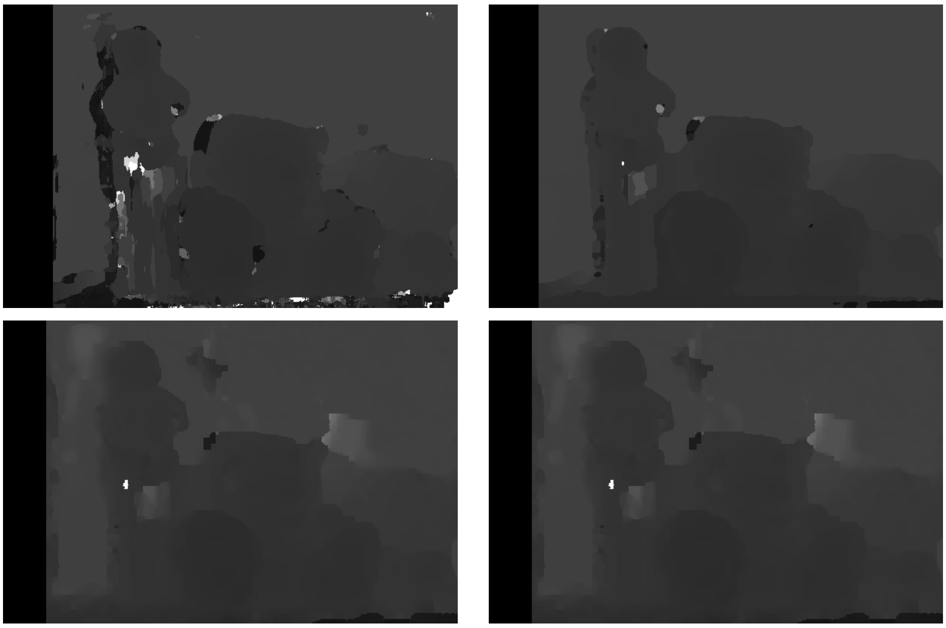


Figure 4.3: Examples of depth maps estimated by FW (up-left), SGM (up-right) and LBP (bottom-left) stereo systems and ground truth depth-map acquired by space-time stereo (bottom-right).

stereo algorithms and due only to occlusions in the case of space-time stereo.

The accuracy A and the precision P of the two systems were computed according to Equation (4.1) and (4.2) respectively using the space-time stereo data as ground truth and are shown in Table 4.1 together with the resolution characteristics.

Table 4.2 reports the execution times of the considered stereo algorithms.

It is worth reminding that the presented results apply to the considered reference scene and not to general scenes. Nevertheless they allow for some concrete and reasonable considerations of general kind based on quantitative data. Namely ToF cameras are typically faster, more accurate and precise than stereo algorithms (with respect to the considered implementations). On the other side, stereo vision systems have better spatial resolution. Stereo depth resolution can be better than ToF resolution for closer objects. ToF cameras depth resolution is less dependent on the object distance than the one of stereo systems. In the case of stereo vision it is possible to change the baseline and focal in order to improve the resolution. The execution times of CPU and GPU stereo algorithms do not allow to obtain frame rates as high as those of ToF cameras. Such complementary characteristics of the two systems open the way to the idea of

Quantity	Stereo FW	Stereo SGM	St LBP	ToF (MESA SR4000)
Accuracy	$60[mm]$	$35[mm]$	$41[mm]$	$25[mm]$
Precision	$13[mm]$	$2[mm]$	$12[mm]$	$11[mm]$
Spatial Res.	777×778	777×778	777×778	176×144
Depth Res.	$\frac{z^{*2}}{fb}$	$\frac{z^{*2}}{fb}$	$\frac{z^{*2}}{fb}$	$0.3[mm] < \delta_z < 1[mm]$

Table 4.1: Experimental comparison between the ToF cameras and the stereo vision systems. Accuracy and precision are computed with respect to the scene shown in Figure 4.2. Spatial resolution and depth resolution are characteristic of the considered acquisition system. The considered stereo system has focal $f = 856.3[pxl]$ and baseline $b = 176.8[mm]$.

Stereo algorithm	Execution time [ms]
FW	$\approx 130[ms]$
SGM	$\approx 2400[ms]$
LBP	$\approx 1600[ms]$

Table 4.2: Execution times of the stereo algorithms. FW and SGM are implemented on CPU, while LBP is implemented on GPU. The experiments were run on a machine with a 4 core Intel i7, $3.06[GHz]$ CPU and NVIDIA NVS 3100M GPU.

fusing their data. Given these considerations it is immediate do notice how ToF cameras and stereo systems are complementary in terms of all the considered metrological quantities.

Intuitively it is possible to guess that considering an acquisition system made of a ToF and two cameras a superior capability of acquiring scene depth information can be obtained, with respect to the two subsystems. However, the proper exploitation of data coming from a ToF camera and a stereo system in order to provide a synergic fusion that allows for better accuracy, precision and resolution is a complex problem to be tackled. Several approaches have been proposed in order to solve such a problem. In the next Chapter a method for ToF and stereo data fusion that aim at obtaining the best of the two subsystems is proposed.

5

Fusion of tof and stereo data: probabilistic approach

Given the considerations expressed in the previous chapters, it is immediate to notice how ToF cameras and stereo systems are complementary in terms of all the considered metrological quantities. Intuitively it is possible to state that by their fusion it is possible to obtain a superior capability by acquiring scene depth information with respect to the the two subsystems. However, the proper exploitation of data coming from ToF cameras and stereo systems in order to provide a synergic fusion that allows for better accuracy, precision and resolution is a complex problem. Several approaches have been proposed in order to solve such a problem and the focus of this chapter is to present a novel approach. After a review of the literature regarding ToF and stereo fusion, it is presented a probabilistic method in order to solve the fusion problem. Such method takes advantage of the classical MAP-MRF Bayesian formulation and exploits the specific properties of ToF and stereo data respectively presented in Chapter 2 and 3. The quality of the proposed method is assessed as function of the metrological quantities introduced in Chapter 4.

5.1 Related Work

Since their introduction, matricial Time-of-Flight range cameras (ToF) have attracted a lot of attention. They have been studied as stand-alone cameras and in multi-sensor setups. A very detailed description of their working principle can be found in the original Ph.D. thesis of Robert Lange [73], as well as in [75]. More recent books [45] also describe in detail how these sensors work and how they can be calibrated and used for accurate 3D measurements. A characterization of the performances of ToF cameras can be found in [58, 67, 83]. In [67], numerous effects that influence ToF cameras range measurements are analyzed and described, and a first distribution model of the ToF camera measurement error is presented. This error distribution model regards the ToF camera as a device obtained by integrating multiple single-point ToF devices in a matricial organization. In [58] a qualitative analysis of the influence of scene reflectance on the quality of depth measurements is reported. The first model of the ToF camera error measurements that accounts for scene properties (*i.e.* depth discontinuity and scene reflectance) is instead presented in [42].

For various types of applications it is interesting to consider the possibility of including ToF cameras in a multi-sensor setup. These setups can be made by the combination of one ToF camera with a single color camera as in [50, 51, 57, 99]. Other approaches [42, 48, 58, 71, 81, 98, 102, 103, 104] exploit two color cameras arranged in a stereo rig in order to have two 3D measures, one from the ToF camera and one from the stereo pair, that are then combined together. In [54] four color cameras are used and finally in [69] multiple ToF cameras and multiple color cameras are employed together.

The setup constituted by one ToF camera and a stereo pair is indeed one one of the most intriguing because the two systems have complementary characteristics. A first naïve approach to this problem is the one of [71], in which the depth-map acquired by the ToF and the depth-map acquired by the stereo pair are separately obtained, then are registered on a unique frame (*e.g.* the stereo pair frame) and finally are averaged. The quality of the results obtained by this method is limited, because the errors of the single acquisition systems propagate. Another simple approach is the one of [58], in which the depth map acquired by the ToF is reprojected on the reference image of the stereo pair, it is then interpolated and finally used as initialization for the application of a dynamic programming stereo vision algorithm. The main issue of this method is that if the information from the range sensor is not correct, the dynamic programming algorithm produces severe artifacts. In [99] an alternate approach based on bilateral filtering is proposed in order to build a 3D value of depth probability (cost volume). The method of [99] can also be generalized to the case of two color cameras instead of only one. In order to reduce the computational burden of the iterative bilateral filtering on the cost volume, a hierarchical version of the bilateral filtering method is proposed in [98]. In [98], after up-sampling the depth map acquired by the ToF by a hierarchical application of bilateral filtering, the authors apply a plane-sweeping stereo algorithm to the acquisition volume defined with respect to the ToF reference frame. Finally the depth information acquired from the ToF and from the stereo are fused together by means of a confidence based strategy.

In [42] a completely different method is proposed, based on a probabilistic formulation. The final depth-map is recovered from the one acquired by the ToF and the one estimated by stereo by performing a ML local optimization in order to increase the accuracy of the depth measurements. The main limitations of this algorithm are that the resolution of the final depth-map is the one of the ToF and the lack of a final global optimization step. Another local approach [48] instead uses a locally-consistent framework to combine the measures of the ToF sensor with the data acquired by the color cameras. The method proposed in [103] is instead based on a MAP-MRF Bayesian formulation, inside which a belief propagation algorithm is used in order to optimize a global energy function. This method allows to increase both resolution and accuracy of the depth measurements performed by each single subsystems. A temporal extension of this method is proposed in [102], and an automatic way to set the weights of the ToF and of the stereo measurements is presented in [104]. All the methods of [102, 103, 104] do not exploit a rigorous model for the ToF measurements. Another recent method [81] uses a variational approach in order to combine the two devices.

Concerning the optimization of the global energy functions that are usually obtained from a MAP-MRF approach, classical methods adopted are: Loopy Belief Propagation (LBP) [82], Graph Cuts (GC) [29], Iterated Conditional Modes (ICM) [21], Tree-Reweighted Message Passing (TRW) [94]. A comprehensive analysis and a comparison of such algorithms are presented in [91]. Since usually these methods are are adopted in problems which present a global energy function defined for a finite set of variables (sites) which can take discrete values (site-wise uniform), they are not directly suited for the optimization of the energy function that is derived in this work and an extension of LBP in order to solve the considered optimization problem is therefore proposed in Section 5.7.

5.2 Depth Estimation from Multiple Devices Measurements

The problem estimating three-dimensional geometry (or depth [1]) from ToF and stereo data can be framed inside the more general problem of the depth estimation from heterogeneous data acquired by multiple devices. This class of problems contemplates the presence of a set of N matricial devices $D_1, ..., D_N$ acquiring the scene (typical examples of matricial devices are standard cameras, ToF cameras, light-coding range cameras and stereo vision systems). These devices respectively acquire the depth measures $I_1, ..., I_N$ arranged on a matricial structure, that can be considered as realizations of the random fields $\mathcal{I}_1, ..., \mathcal{I}_N$. The goal of the various approaches for this problem is the estimation of scene depth-map Z, which can be regarded as the realization of a random field $\mathcal{Z}$. The estimated scene depth-map is indicated as $\hat{Z}$ and is generally desired to be characterized by

- an high accuracy, in the sense of small mean depth estimation error
- an high precision, in the sense of high depth measurements repeatability
- an high resolution, in terms of both depth resolution (high sampling of depth values) and spatial resolution (number of points in the depth-map)

The estimate $\hat{Z}$ generally can be calculated within a probabilistic framework as the solution of a Maximum-A-Posteriori (MAP) problem

$$\hat{Z} = \arg\max_{Z \in \mathcal{Z}} P(Z|I_1, ..., I_N) \tag{5.1}$$

in which $P(Z|I_1, ..., I_N)$ is the posterior probability of the scene depth-map Z, given the acquired data $I_1, ..., I_N$. By applying Bayes rule, Equation (5.1) can be rewritten as

$$P(I_1, ..., I_N|Z)P(Z) \tag{5.2}$$

in which $P(I_1, ..., I_N|Z)$ is the likelihood of the measurements $I_1, ..., I_N$ given the scene depth distribution, and $P(Z)$ is the scene depth prior probability. This formulation is interesting because it allows to decouple the properties of the scene ($P(Z)$) with the

[1] In the context of matricial devices, three-dimensional geometry and depth information are equivalent concepts. In this chapter the two notations are used indifferently.

measurement characteristics of the sensors $D_1, ..., D_N$ ($P(I_1, ..., I_N|Z)$). The problem of Equation (5.2) can be very complex, since the relationships between the measurement errors of the various sensors are complicated and hard to model. It is quite common to suppose that the measurement errors of the various sensors are independent, obtaining therefore the easier problem formulation

$$\hat{Z} = \arg\max_{Z \in \mathcal{Z}} P(I_1|Z)...P(I_N|Z)P(Z) \tag{5.3}$$

in which $P(I_n|Z), n = 1, ..., N$ are the likelihood probabilities (or simply *likelihoods*) of the single device measurements given the scene depth Z. This hypothesis of independence has been adopted in [62] for the case in which there are two stereo vision systems acquiring the scene from two different points-of-view and in [42, 102, 103, 104] in the case of two sensors, being a stereo vision system and a ToF camera. Especially in this second situation, which is also the situation considered through this chapter, the independence assumption seems very likely, because the measurement errors of ToF cameras and stereo vision systems are pretty different. The main intuition behind this independence assumption is related to the fact that the sources of error for stereo and ToF measurements are very different. Errors of ToF cameras are influenced by scene illumination at the radiating IR wavelength and by scene reflectivity at such wavelength, while errors of stereo vision systems are related to the amount of texture in the scene. The problem formulation reported in Equation (5.3) addresses a very general problem in a simple and tractable form, emphasizing the different components that play a role in the problem. Such components are

1. The likelihood of the single device measurements $P(I_n|Z), n = 1, ..., N$. These quantities need to be modeled very carefully, accounting for the theoretical principles behind the adopted sensors.

2. The scene depth prior probability (or simply *prior*) $P(Z)$, which has to take in account for the for the nature of the acquired scene. It can have a very general form in the case in which the scene is generic, but it can be also very specific if the acquired scene has some peculiar and well-known characteristics.

3. The maximization of the probability terms, which has to take in account for the properties of the likelihood and prior terms, and exploit them in order to be effective and efficient.

This work improves the state of the art for points 1) and 3), while adopting a common approach to point 2).

Concerning the first point, a formal model of the ToF and stereo likelihood is proposed. Especially in the case of the ToF cameras, this is the first time in which a formal model is exploited in this context, and this is a major advancement that this thesis provides. In fact, one of the main limitations of previous works [102, 103, 104] is that they are characterized by the exploitation of a simple heuristic model for ToF cameras, even through a formal model of their likelihood is available [45, 73]. In the proposed approach, not only the formal model described in Chapter 2 is adopted, but it is also extended in order to remove some of it limitations.

Concerning the third point, a specific maximization algorithm based on Loopy-Belief-Propagation (LBP) is adopted. Such algorithm exploits the nature of the quantities to be optimized, and it shows to be more efficient than the classical LBP adopted in other approaches [102, 103, 104]. In fact, the amount of operations performed by the proposed optimization algorithm is just 7% of the amount of operations performed by the classical approach. Moreover the proposed algorithm generally produces more accurate depth estimates.

5.2.1 Depth Estimation from ToF and Stereo Data

In this specific case two acquisition sensors are considered, namely a ToF camera T and a stereo vision system S (also called stereo setup). The stereo setup is constituted by a couple of color cameras L and R, respectively the left and the right camera. The proposed framework does not require a specific position arrangement of the three devices. However in the setup used for the experimental results (shown in Figure 5.1) the ToF camera T is placed in between the two cameras L and R for symmetry purposes. Associated to T there are a standard 3D CCS, called T-3D CCS, with axes $\{\mathbf{x}_T, \mathbf{y}_T, \mathbf{z}_T\}$, and a standard 2D CCS, called T-2D CCS, with axes $\{\mathbf{u}_T, \mathbf{v}_T\}$. Associated to L and R there are a couple of standard 3D CCSs, called L-3D and R-3D CCS, with axes $\{\mathbf{x}_L, \mathbf{y}_L, \mathbf{z}_L\}$ and $\{\mathbf{x}_R, \mathbf{y}_R, \mathbf{z}_R\}$ respectively, and a couple of standard 2D CCS, called L-2D and R-2D CCS, with axes $\{\mathbf{u}_L, \mathbf{v}_L\}$ and $\{\mathbf{u}_R, \mathbf{v}_R\}$. The dispositions of the various CCSs is reported in Figure 5.2. As it is described in the next section, a TOF camera acquires the following data at video-rate:

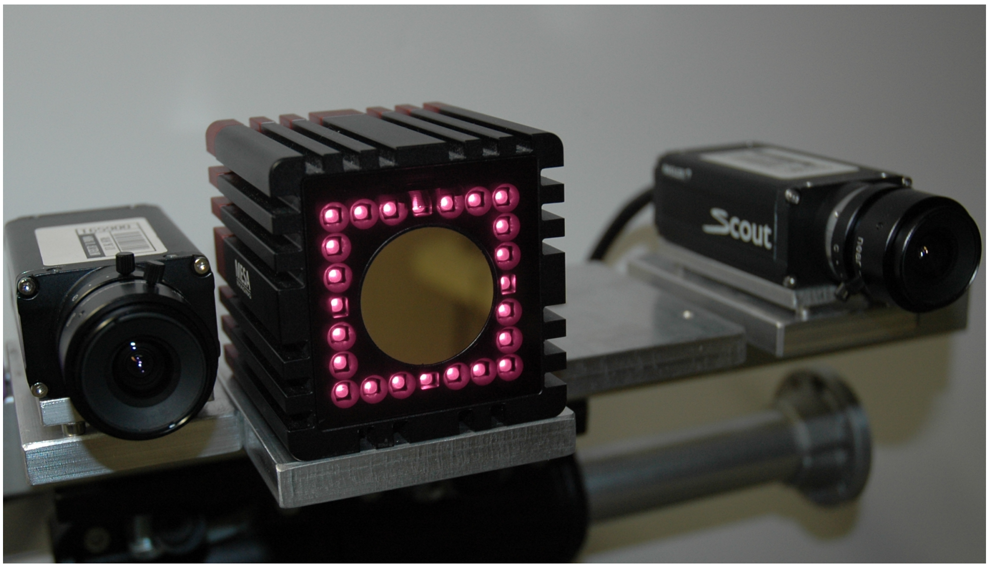


Figure 5.1: Considered acquisition system constituted by a ToF camera T and by a stereo vision system $S \triangleq \{L, R\}$.

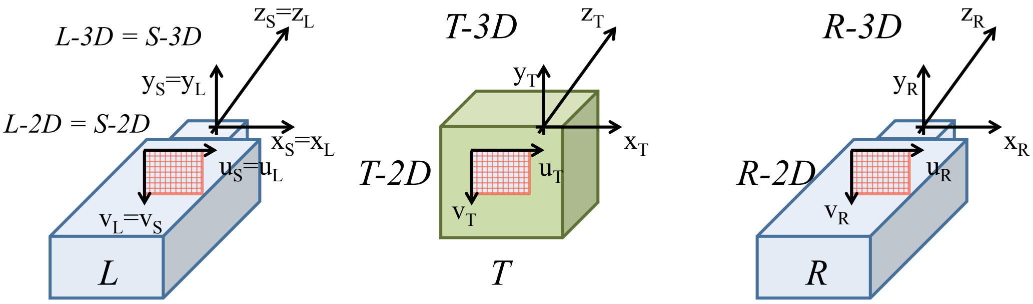


Figure 5.2: CCSs (3D and 2D) associated to the various sensors constituting the acquisition system.

1. an amplitude image A_T, defined on the lattice Λ_T associated to the T-2D CCS, which can be considered as the realization of a random field $\mathcal{A}_T$.

2. an intensity image B_T, defined on the lattice Λ_T, which can be considered as the realization of a random field $\mathcal{B}_T$.

3. a depth-map Z_T, defined on the lattice Λ_T, which can be considered as the realization of a random field $\mathcal{Z}_T$.

In this chapter the ensemble of such data are usually indicated as $I_T \triangleq \{A_T, B_T, Z_T\}$.

The data acquired by the two cameras L and R are synchronized pairs of color images addressed as I_L and I_R respectively. Images I_L are defined on the lattice Λ_L associated to the L-2D reference frame, and can be considered as the realization of a random field $\mathcal{I}_L$. Images I_R are defined on the lattice Λ_R associated to the R-2D reference frame, and can be considered as the realization of a random field $\mathcal{I}_R$. Data acquired by S are also denoted as $I_S \triangleq \{I_L, I_R\}$. The CCSs of L are considered also as reference for stereo data. An example of data acquired by T and S is reported respectively in Figure 5.3 and 5.4.

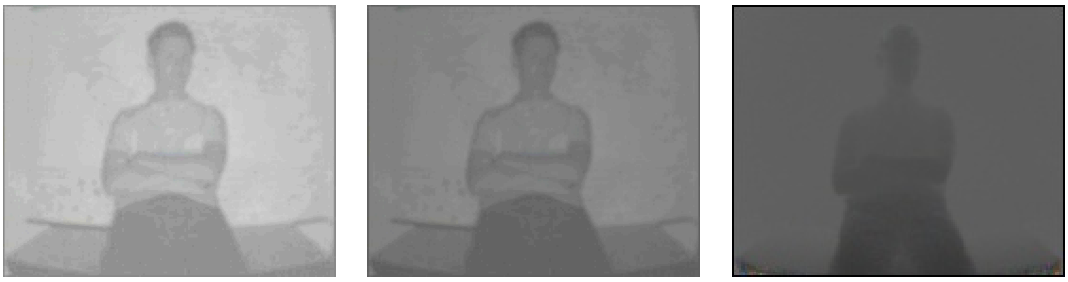


Figure 5.3: Data acquired by T: A_T (left), B_T (center) and Z_T (right). Images A_T and B_T have been manipulated in order to increase visibility on printed paper.

Given this notation, it is possible to express Equation (5.3) for the specific case of an acquisition system constituted by a ToF camera T and a stereo S as

$$\hat{Z} = \arg\max_{Z \in \mathcal{Z}} P(I_T|Z)P(I_S|Z)P(Z) \tag{5.4}$$

in which $P(I_T|Z)$ is the likelihood of ToF measurements given the scene depth, $P(I_S|Z)$ is the likelihood of stereo measurements given the scene depth, and $P(Z)$ is the scene depth prior probability. The various components of Equation (5.4) are analyzed and

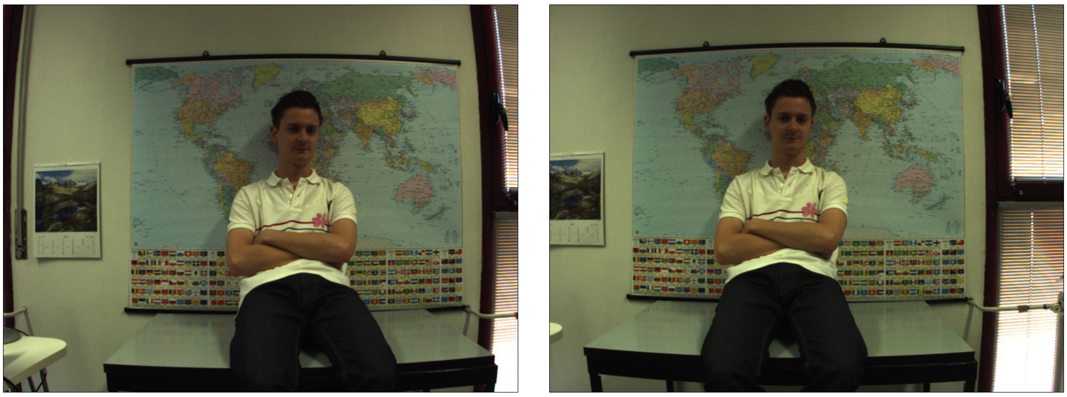


Figure 5.4: Data acquired by S: I_L (left) and I_R (right).

described in the following sections. Let us finally notice how in Equation 5.4 there are input quantities I_T and I_S defined with respect to different CCSs (*i.e.*, the T-CCSs for I_T and the S-CCSs for I_S). Such quantities need to be referred with respect to a unique CCS. To this purpose it is necessary to jointly calibrate T and S. Such calibration is performed with the procedure introduced in [42].

5.3 ToF Likelihood

As presented in Chapter 2, the distribution of the depth acquisition noise of a ToF pixel can be approximated as a Gaussian with standard deviation

$$\sigma_\rho = \frac{c}{4\pi f_{mod}\sqrt{2}}\frac{\sqrt{B}}{A} \tag{5.5}$$

Standard deviation (5.5) determines the precision (repeatability) of the distance measurement and it is directly related to f_{mod}, A and B. The model of Equation (5.5) although is a well-known theoretical model in the fields of ToF design and metrology has never been exploited in computer vision problems. One of the main limitations of such model is that is does not take into account practical issues that arise when dealing with actual ToF cameras, such as the finite size of sensor pixels. In order to account for such non ideality we propose a generalized version of Equation (5.5), obtaining a more realistic model suitable for the construction of a reliable likelihood of the ToF depth measurements $P(I_T|Z)$. Let us consider a point p_i in the lattice Λ_T. As shown in Figure 5.5, p_i is relative to a sensor pixel of finite size which acquires information

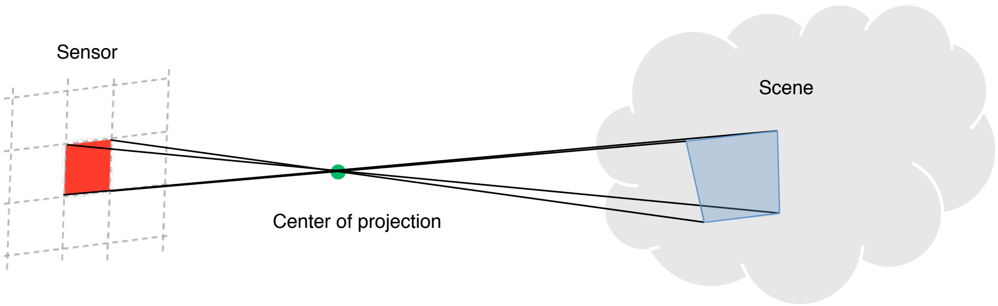


Figure 5.5: Example of $p_i \in \Lambda_T$ relative to a finite size sensor pixel that is associated to a scene area of finite size as well.

relative to a scene area of finite size as well. If the finite scene area is flat, than the first order Taylor approximation of the scene area with a fronto-parallel plane is realistic and therefore the model of Equation (5.5) is still valid. However, if there is a depth discontinuity in the finite scene area, the first order Taylor approximation is not correct, and therefore the model of Equation (5.5) is not valid. In particular, let us consider the case of a scene area associated with p_i constituted by two different regions R_C (closest region) and R_F (furthest region) divided by a depth discontinuity. The region R_C is approximately at depth z_C, and the region R_F is approximately at depth z_F. The depth measured by the pixel associated to the point p_i is

$$\tilde{z}_i = \alpha z_C + (1 - \alpha) z_F \tag{5.6}$$

in which α is the percentage of scene area associated to R_C and consequently $(1 - \alpha)$ is the percentage of scene area associated to R_F. In order to obtain a likelihood of the T depth measurements of z_i it is necessary to understand which values of the depth z_i of p_i are most likely if there is a measurement $\tilde{z}_i$. As shown in Figure 5.6 it is possible to make a distinction between two situations

1. if R_C and R_F belong to two different surfaces, the actual depth might be close to either z_C or z_F, and not somehow in between the two distances (Figure 5.6.a). This situation can be called *disconnected discontinuity.*

2. if R_C and R_F belong to the same surface the actual depth might can be either close to z_C or z_F or somewhere in between the two (Figure 5.6.b). This situation can be called *connected discontinuity.*

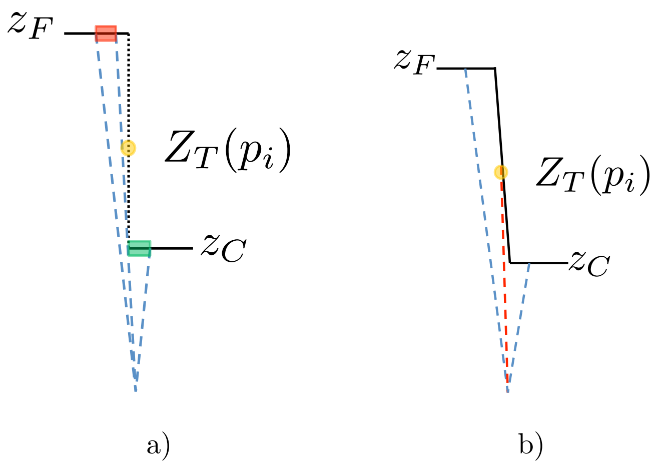


Figure 5.6: Disconnected discontinuity (a) and connected discontinuity (b).

It is not known a priori which situation occurs, hence it is necessary to provide a model that accounts at the same time for each of the two scenarios. In order to come to such a model it is worth exploiting the fact that if p_i is relative to a scene area crossed by a discontinuity between R_C and R_F, some of the points p_j in the 8-neighborhood $\mathcal{N}(p_i)$ of p_i are relative to points at distance z_C, and some others to points at distance z_F. This concept is illustrated in Figure 5.7 It is therefore possible to exploit this intuition in order to obtain a likelihood term for p_i accounting for the fact that if p_i is across a discontinuity the actual value of p_i might be around the one measured by the pixel relative to p_i itself (connected discontinuity) or around some of the depth measurements relative to its 8-neighbors p^i_j (connected or disconnected discontinuity). The fusion of contributions from neighboring pixels can be done by considering a classical image correlation model [97], obtaining therefore the following expression of the ToF likelihood for the point p_i

$$P(I_T|Z) \propto \mathcal{N}(z_i, \sigma_i^2) + e^{-1}\sum_{j=1}^{4}\mathcal{N}(z^i_j, \sigma_{ij}^2) + e^{-2}\sum_{j=5}^{8}\mathcal{N}(z^i_j, \sigma_{ij}^2) \tag{5.7}$$

in which $z^i_j = z(p^i_j)$, and σ_i and σ_{ij} are the standard deviations of the depth measurements for the points p_i and p^i_j respectively, obtained according to Equation (5.5). The explicit version of Equation (5.7) is

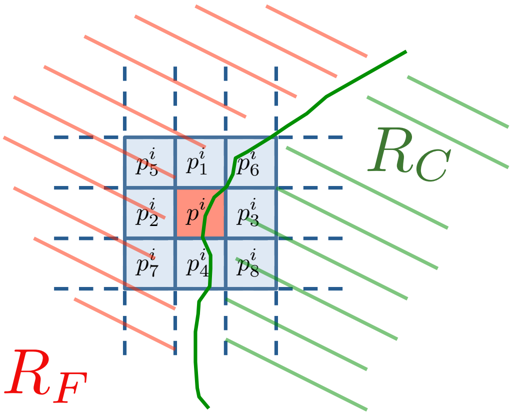


Figure 5.7: Discontinuity between R_C and R_F crosses the area associated to p_i, p_6^i, p_6^i. points $p_1^i, p_2^i, p_5^i, p_7^i$ are in the same scene region of R_F while points p_3^i, p_8^i are in the same region of R_C.

$$
\begin{aligned}
P(I_T|Z) \quad & \propto \quad \frac{1}{\frac{c}{4\pi f_{mod}\sqrt{2}}\frac{\sqrt{B(p_i)}}{A(p_i)}} \exp -\left(\frac{z-z_i}{\frac{c}{4\pi f_{mod}\sqrt{2}}\frac{\sqrt{B(p_i)}}{A(p_i)}}\right)^2 \\
& + \quad e^{-1}\sum_{j=1}^{4} \frac{1}{\frac{c}{4\pi f_{mod}\sqrt{2}}\frac{\sqrt{B(p_j^i)}}{A(p_j^i)}} \exp -\left(\frac{z-z_j^i}{\frac{c}{4\pi f_{mod}\sqrt{2}}\frac{\sqrt{B(p_j^i)}}{A(p_j^i)}}\right)^2 \\
& + \quad e^{-2}\sum_{j=5}^{8} \frac{1}{\frac{c}{4\pi f_{mod}\sqrt{2}}\frac{\sqrt{B(p_j^i)}}{A(p_j^i)}} \exp -\left(\frac{z-z_j^i}{\frac{c}{4\pi f_{mod}\sqrt{2}}\frac{\sqrt{B(p_j^i)}}{A(p_j^i)}}\right)
\end{aligned}
\tag{5.8}
$$

Before moving forward, let us analyze what Equation (5.7) and (5.8) mean and why the proposed model for the ToF likelihood is adequate. Let us start by saying that if there is not a depth discontinuity, the various Gaussian contributions have similar mean, and therefore the ToF likelihood becomes very similar to Equation (5.5). Therefore this model, although being more general than the one of Equation (5.5), reduces to such a model in the particular case in which its assumptions are valid. In case of a discontinuity, the model of (5.7) and (5.8) is likely to assign an high probability to distances around z_C, around z_F and around the measured distance z_i, contemplating therefore both the two cases of Figure 5.7.

The fact that all the terms in (5.7) and (5.8) are Gaussians leads to nice properties based on the concept of *useful interval*. In fact it is worth to notice that, given certain depth measurements for a pixel and its neighborhood, it is likely that the actual depth value z^* is not very different from at least one of them. It is possible to formalize this concept by noticing that the likelihood of (5.7) and (5.8) is a mixture of Gaussians. For a Gaussian distribution the concept of *useful interval* ensures that with probability 0.997 the actual value of the measured quantity stays in the interval $[\mu - 3\sigma, \mu + 3\sigma]$ where μ is the mean and σ is the standard deviation of the Gaussian distribution. In the case of a mixture of Gaussians the useful interval can be defined as $[\mu_{min} - 3\sigma_{min,\mu_{max}+3\sigma_{max}}]$, in which μ_{min} and σ_{min} are the mean and the standard deviation of the Gaussian in the mixture with minimum mean value and σ_{max} are the mean and the standard deviation of the Gaussian in the mixture with maximum mean value. In the specific case of depth measurements, μ_{min} and σ_{min} can be named μ_C and σ_C (where C stays for "close") and μ_{max} and σ_{max} can be named μ_F and σ_F (where F stays for furthest). All the possible depth values for the specific pixel which are not in its associated useful interval can simply be ignored. This concept allows to prevent useless computations in the fusion algorithm as it explained in the following sections. An example of useful interval for a pixel is reported in Figure 5.8.

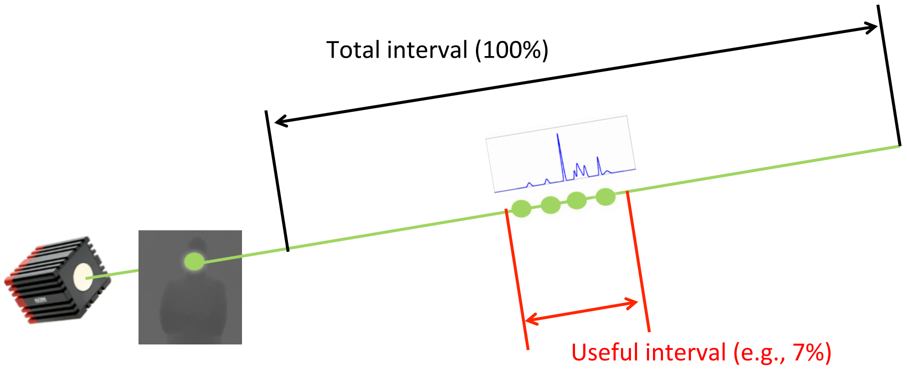


Figure 5.8: The useful interval concept might allow for a computational reduction. In particular, with respect to the experimental results of Section 5.8 such reduction allows to perform 7% of the computations that should be done by considering the total interval

.

The sampling of depth values inside the *useful interval* influences the allowed quality of depth resolution. From a higher-level point of view it is possible to say that the model of (5.7) and (5.8) accounts for both the theory behind classical ToF measurement error distributions and for the matricial nature of ToF cameras sensors. As it is shown in Section 5.8, this model is able to lead to accurate depth estimation.

5.4 Stereo Likelihood

Several approaches have been considered in order to solve the problem of modeling stereo likelihoods for the case of a calibrated and rectified stereo vision system [28, 89]. The general idea behind the various proposed methods is that, similarly to the case of stereo (Chapter 3), given a certain scene depth distribution Z it is possible to identify a set of couples of points $(p_i^L, p_i^R), p_i^L \in \Lambda_L, p_i^R \in \Lambda_R$ such that they refer to a unique 3D point P_i in the scene. Points p_i^L and p_i^R are called *conjugate points* (ore simply *conjugates*). The coordinates of p_i^L are $\mathbf{p}_i^L = [u_i^L, v_i^L]^T$ and the coordinates of p_i^R are $\mathbf{p}_i^R = [u_i^R, v_i^R]^T$. If the stereo system has undergone rectification, points p_i^L and p_i^R share the same vertical coordinate ($v_i^L = v_i^R$) while their horizontal displacement (disparity $d_i = u_i^L - u_i^R$) is proportional to the inverse of the depth z_i of P_i (3.1). The likelihood of stereo data given the depth distribution z_i can be obtained by considering multiple hypothesis $z_{i,n}, n = 1, ..., N$ for the depth z_i and computing a likelihood value for each of such hypotheses. In this way a likelihood distribution can be obtained. The likelihood distribution $P(I_S|Z(P_{i,n}))$ for hypotheses $z_{i,n}, n = 1, ..., N$ can practically be computed as follows by taking advantage of classical stereo schemes (Chapter 3)

1. for each depth hypothesis $z_{i,n}, n = 1, ..., N$ compute the 3D coordinates of the corresponding 3D point $P_{i,n}$

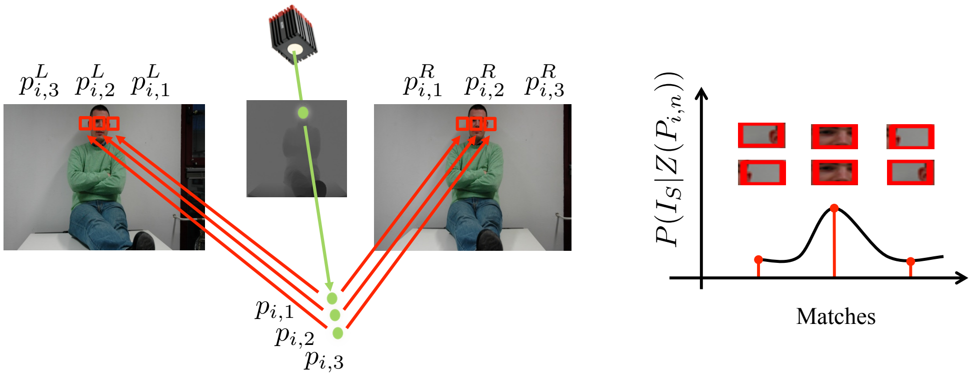
2. project $P_{i,n}$ into the 2D points $p_{i,n}^L \in \Lambda_L, p_{i'n}^R \in \Lambda_R$. The coordinates of $p_{i,n}^L$ are $\mathbf{p}_{i,n}^L = [u_{i,n}^L, v_{i,n}^L]^T$ and the coordinates of $p_{i,n}^R$ are $\mathbf{p}_{i,n}^R = [u_{i,n}^R, v_{i,n}^R]^T$.

3. consider a window $W_{i,n}^L$ centered around $p_{i,n}^L$ and the window $W_{i,n}^R$ centered around $p_{i,n}^R$

4. evaluate the similarity (hence the likelihood) between $I_L(W_i^L)$ and $I_R(W_i^R)$

An example of this procedure is reported in Figure 5.9.

What is still missing at this point is the actual computation of the similarity between $I_L(W_i^L)$ and $I_R(W_i^R)$. A classical method for computing such similarity for


$p^L_{i,3}$ $p^L_{i,2}$ $p^L_{i,1}$
$p^R_{i,1}$ $p^R_{i,2}$ $p^R_{i,3}$
$p_{i,1}$
$p_{i,2}$
$p_{i,3}$
$P(I_S|Z(P_{i,n}))$
Matches

a) b)

Figure 5.9: Example of stereo likelihood calculation. The 3D points sampled from the tot useful interval are re-projected onto the two stereo images (a) and the stereo likelihood is computed by matching the windows centered on the conjugate couples (b).

likelihood modeling purposes is the one proposed in [89]. Such method is based on the cost matching calculated according to the Birchfield-Tomasi method [23] without imposing any cost aggregation procedure (a definition and an exhaustive analysis of cost matching and cost aggregation procedures is reported in [86]). Recently some advancements in stereo cost aggregation procedures have shown that it is possible to obtain improved results by accounting for image segmentation clues [77]. On the light of such recent advancements we propose a method for calculating the likelihood of stereo measurements that improves the framework of [23] by accounting also for the method of [77]. Differently from [89], we adopt Truncated Absolute Difference (TAD) for the matching cost computation as in [77] instead of the Birchfield-Tomasi method. Let us assume that the segmentations S_L and S_R of images I_L and I_R respectively are obtained by applying an image segmentation method (*e.g.*, the one of [3]) and that $W^L_{i,n}$ and $W^R_{i,n}$ are rectangular windows of size $(2H_W + 1) \times (2W_W + 1)$, centered at $p^L_{i,n}$ and $p^R_{i,n}$ respectively. The likelihood of stereo measurements $P(I_S|Z(P_{i,n}))$ can be calculated as

$$P(I_S|Z(P_{i,n})) = \frac{\exp -\frac{\mathcal{C}(\mathbf{p}^L_{i,n},\mathbf{p}^R_{i,n})}{\sigma^2_I}}{\sum_{k=1}^{N} \exp -\frac{\mathcal{C}(\mathbf{p}^L_{i,k},\mathbf{p}^R_{i,k})}{\sigma^2_I}} \tag{5.9}$$

in which $\mathcal{C}(\mathbf{p}^L_{i,n}, \mathbf{p}^R_{i,n})$ (and similarly $\mathcal{C}(\mathbf{p}^L_{i,k}, \mathbf{p}^R_{i,k})$) is computed as

$$
\begin{aligned}
\mathcal{C}(\mathbf{p}^L_{i,n}, \mathbf{p}^R_{i,n}) \quad = \quad & \tfrac{1}{(2H_W+1)\times(2W_W+1)} * \\
& \textstyle\sum_{u\in[-W_W,W_W]} \sum_{v\in[-H_W,H_W]} \\
\{ \quad & w(\mathbf{p}^L_{i,n}, [u^L_{i,n} - u, v^L_{i,n} - v]^T) * \\
& w(\mathbf{p}^R_{i,n}, [u^R_{i,n} - u, v^R_{i,n} - v]^T) * \\
& \min\left(I_L(\mathbf{p}^L_{i,n}) \ominus I_R(\mathbf{p}^R_{i,n}), T_h\right) \quad \}
\end{aligned}
\tag{5.10}
$$

where T_h is the TAD threshold parameter, $\ominus$ is the operator defined as the geometric mean of the three intra-channel difference between I_L and I_R and $w(\mathbf{p}, \mathbf{q})$, with $\mathbf{p} = [u_p, v_p]^T, \mathbf{q} = [q_p, q_p]^T$ is the aggregation weight of [77], calculated as

$$
w(\mathbf{p}, \mathbf{q}) \triangleq \begin{cases} 1 & if S() == S() \\ I() \ominus I() & otherwise \end{cases}
\tag{5.11}
$$

in which S is the segmented image on which p and q belong (either S_L or S_R) and I is the acquired color image (either I_L or I_R).

5.5 Scene Depth Prior

Other elements of Equation (5.4) that are still not described are the choice of the lattice Λ_Z on which the output depth-map is defined and the characteristics of the prior probability of the scene depth $P(Z)$. The choice of Λ_Z, *i.e.*, the lattice on which the final scene depth-map is estimated, is a very important task that strongly characterizes the performances of the fusion algorithm both in terms of computation resources and results precision. Since the trinocular system is constituted by a stereo system S and a ToF camera T, in literature have been presented two different choices of Λ_Z: $\Lambda_Z \equiv \Lambda_S$ or $\Lambda_Z \equiv \Lambda_T$. The first choice has been adopted by [58, 71, 99, 102, 103], while the second by [42, 98].

On one hand, the choice of considering $\Lambda_Z \equiv \Lambda_S$ allows to adopt a standard expression of the stereo likelihood, as the ones proposed in [28, 89]. In this case, the ToF likelihood can be expressed only in heuristic way, as proposed in [102, 103]. The main advantages of this method are:

- the adopted stereo likelihood model is well consolidated

- the final resolution of $\hat{Z}$ is the high resolution of the stereo pair images $\{I_L, I_R\}$
- the ToF and stereo data fusion problem is re-conduced to an extended version of classical stereo vision algorithms, and therefore it is possible to exploit all the powerful tools that are currently available for the solution of the stereo vision problem, *e.g.*, Graph Cuts [28] and Belief Propagation [89].

The main disadvantages of this method are:

- the proposed fusion frameworks do not exploit the availability of a formal model for the ToF camera, but adopt only heuristic models (*e.g.*, the ToF camera error model proposed in [102, 103])
- the fusion framework is computationally overwhelming, as it will be explained later
- the final results are not very accurate, since in order to limit the computational complexity several approximations are imposed

On the other hand, the choice of adopting $\Lambda_Z \equiv \Lambda_T$ allows for the exploitation of the previously introduced formal model for both the ToF and the stereo likelihoods, leading to a computationally lighter framework that is also able to deliver more accurate results. This intuitive reasoning is clarified in the following. The greatest disadvantage of this method is that the final lateral resolution of the estimated depth-map $\hat{Z}$ is the same of the data acquired by T, that generally are characterized by a low resolution (*e.g.*, 176 × 144 for the MESA SR4000 [8]).

From the analysis of the two solutions, it is possible to notice how their features are complementary. It would be interesting to have a choice of Λ which leads to an estimated depth-map $\hat{Z}$ characterized by high resolution ($\|\Lambda_Z\| \approx \|\Lambda_L\|$) and that exploit a formally-defined likelihood of the ToF in order to design a light and accurate fusion framework. In order to obtain such a Λ_Z, we decided to adopt an interpolated by L times version of Λ_T: Λ_T^L for which:

- the estimated depth-map $\hat{Z}$ is characterized by an high resolution that is $L\|\Lambda_T\| \approx \|\Lambda_L\|$ (*e.g.*, $L = 2, 4, 6$)
- the ToF likelihood is obtained in a formal way according to the method proposed in Section 5.3

- the stereo likelihood is expressed according to the method proposed in Section 5.4

While the adoption of Λ_T^L in the case of the stereo system does not introduce any structural modification, in the case of a ToF it is necessary to revisit the way in which $P(I_T|Z)$ is calculated.

Let us recall that D_T, A_T and B_T are defined on the low resolution lattice Λ_T. In order to obtain an output depth-map characterized by high resolution it is necessary to up-sample the likelihood $P(I_T|Z)$ from the lattice Λ_T to Λ_T^L. Being L an integer, Λ_T is a sub-lattice of Λ_T^L. We propose to perform a bilinear interpolation of the likelihood probability. Since the concept of spatial-interpolation of probabilities (considered as the process of obtain a "backward-compatible" probability function defined on a up-sampled lattice from a probability function defined on a low-resolution lattice) is not the same of the spatial-interpolation of images or depth-maps, we preferred to adopt a "bilinear interpolation" model, which naturally relates to standard correlation models for 2D random fields. More complex models, such as bicubic interpolation might also be considered for this task. After such an interpolation, it is available an up-sampled likelihood probability distribution of the measurements performed by the ToF camera T: $P^L(I_T|Z)$. For presentation purposes, the superscript L will be omitted from the previous notation.

Let us recall that the random field $\mathcal{Z}$ is defined on the lattice Λ_Z. For each point $p_i \in \Lambda_Z$ there are N_i possible distances $z(p_i^n), n = 1, ..., N_i$. For a specific realization Z of $\mathcal{Z}$, characterized by the per-pixel values: $Z(p_i) = z_i^{n_i}, n_i \in [1, ..., N_i]$, the probability density function is $P(\mathcal{Z} = Z)$. If $\mathcal{Z}$ is assumed to be a Markov Random Field (MRF), the pdf $P(\mathcal{Z} = Z)$ can be expressed as:

$$P(\mathcal{Z}(p_i) = z_i^{n_i}|\mathcal{Z}(p_j) = z_j^{n_j}) \tag{5.12}$$

where $p_j \in \mathcal{N}(p_i)$, being $\mathcal{N}(p_i)$ the neighborhood of p_i, $p_i \in \Lambda_Z$,$n_i \in [1, ..., N_i]$ and $n_j =\in [1, ..., N_j]$. We adopted the classical first order neighborhood $\mathcal{N}_1$(4-neigborhood) as $\mathcal{N}$. Since $\mathcal{Z}$ is a MRF, it is possible to apply the Hammersley-Clifford Theorem [76], and therefore $\mathcal{Z}$ is characterized by a Gibbs distribution, that can be expressed as:

$$P(\mathcal{Z}(p_i) = Z_i^{n_i}|\mathcal{Z}(p_j) = z_j^{n_j}, j : z_j \in \mathcal{N}(p_i)) = \frac{1}{\mathscr{Z}} \exp - \frac{U(z_i^{n_i}; \{z_j^{n_j}, j : p_j \in \mathcal{N}(p_i)\})}{T} \tag{5.13}$$

where $p_i \in \Lambda_Z$, $\mathscr{Z}$ is the so-called partition function, T is the so-called ambient-temperature and $U(z_i^{n_i}; \{z_j^{n_j}, j : p_j \in \mathcal{N}(p_i)\}))$ is the energy of the Gibbs distribution evaluated in the point p_i and in its 4-neighbors $p_j \in \mathcal{N}(p_i)$. The energy of the Gibbs distribution is defined as the sum among all the various cliques of a potential function $V(z_i^{n_i}, z_j^{n_j})$:

$$U(z_i^{n_i}; \{z_j^{n_j}, j : p_j \in \mathcal{N}(p_i)\}) \triangleq \sum_{j:p_j \in \mathcal{N}(p_i)} V(z_i^{n_i}, z_j^{n_j}) \tag{5.14}$$

Concerning the choice of the potential function $V(z_i^{n_i}, z_j^{n_j})$, it is adopted the classical truncated quadratic function $V(z_i^{n_i}, z_j^{n_j}) \triangleq \min\left((z_i^{n_i} - z_j^{n_j})^2, Th\right)$, where Th is the threshold that allows the potential function to be regarded as a robust estimator, as showed in [25]. The final expression for the prior $P(\mathcal{Z} = Z)$ takes the form:

$$P(\mathcal{Z}(p_i) = z_i^{n_i} | \{\mathcal{Z}(p_j) = z_j^{n_j}, j : p_j \in \mathcal{N}(p_i)\}) = \frac{1}{\mathscr{Z}} \exp - \frac{\sum_{j:p_j \in \mathcal{N}(p_i)} V(z_i^{n_i}, z_j^{n_j})}{T} \tag{5.15}$$

Let us recall that such a prior probability is only the last term in the RHS of Equation (5.4). The other two terms are the likelihood probabilities of the ToF $P(I_T|Z)$ and of the stereo $P(I_S|Z)$. As shown in Section 5.3 and 5.4 respectively, the two likelihoods can be evaluated per-pixel, since there is a per-pixel independence. As previously said, $\mathcal{Z}$ is a MRF. Therefore it is possible to state that the combination of the two, *i.e.*, the posterior probability distribution respects the Markovian property. Therefore, the posterior probability can be finally expressed as

$$\begin{aligned} P(\mathcal{Z}(p_i) = z_i^{n_i} | I_S, I_T) \quad &= \quad P(I_S | \mathcal{Z}(p_i) = z_i^{n_i}) P(I_T | \mathcal{Z}(p_i) = z_i^{n_i}) \quad * \\ & \qquad P(\mathcal{Z}(p_i) = z_i^{n_i} | \{\mathcal{Z}(p_j) = z_j^{n_j}, j : p_j \in \mathcal{N}(p_i)\}) \end{aligned} \tag{5.16}$$

where $P(I_S|\mathcal{Z}(p_i) = z_i^{n_i})P(I_T|\mathcal{Z}(p_i) = z_i^{n_i})$ is the relative in the specific problem of the so-called *data term* P_{data} and $P(\mathcal{Z}(p_i) = z_i^{n_i}|\{\mathcal{Z}(p_j) = z_j^{n_j}, j : p_j \in \mathcal{N}(p_i)\})$ is the *smoothness term* P_{smooth}. The data term assures that the probability of the depth distribution given the measurements is defined on the basis of the measurements themselves, and the smoothness term imposes the piecewise-smoothness of the estimated scene surface.

5.6 Building Posterior

Now that all the optimization terms of Equation (5.4) are defined it is possible to explain how the posterior probability is calculated as follow

- As first step the ToF likelihood described in Equation (5.7) and (5.8) is computed for each point of Λ_Z defined as the interpolated version of Λ_T by L times.
- For each pixel-ToF-likelihood it is possible to consider the relative useful interval and it is possible to sample such interval, as shown in Figure 5.9. As already said, the sampling of this interval determines the allowed depth resolution of the fusion algorithm. Dense sampling leads to good depth resolution and coarse sampling leads to poor depth resolution.
- Each of the sampled points can be projected onto the left and right stereo images and a stereo likelihood can be computed according to Equation (5.9) and 5.10).
- It is not granted that the projection of a point into the stereo images has integral coordinates. Hence stereo images have to be resampled in order to compute the quantities of (5.7) and (5.8). Such resampling is performed via bicubic interpolation.
- The values of the ToF and of the stereo likelihoods are multiplied, obtaining the joint likelihood of the measurements.
- The final outcome of such operations is a set of depth values for each point in Λ_Z with associated a measurement likelihood. The sets of depth values for different points in Λ_Z are different at each location, as shown in Figure 5.10.
- For each point p_i in Λ_Z and for each point $p_j \in \mathcal{N}(p_i)$ it is possible to compute the scene depth prior distribution for the various depth configurations as described in Section 5.5.
- By considering together the joint likelihood and the scene prior it is possible to obtain an explicit value of the posterior probability distribution in the LHS of Equation (5.4), obtaining therefore a global energy function to be optimized.

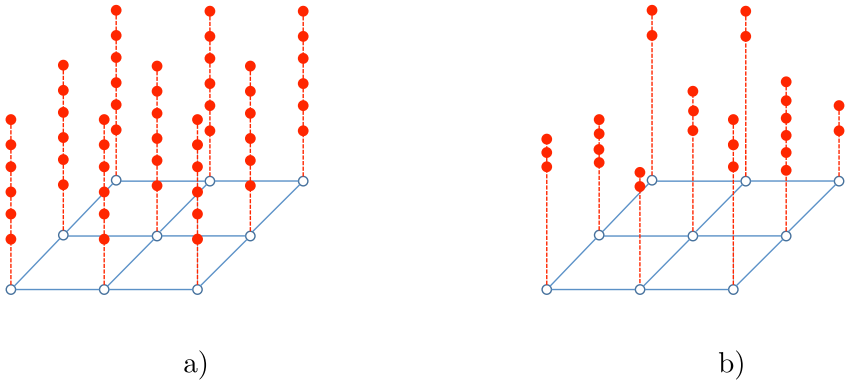


Figure 5.10: Differently from classical problems (a) in this particular problem a set of different values of the range is associated to each point in the domain Λ_Z (b).

5.7 Loopy-Belief-Propagation optimization

Concerning the optimization of the global energy functions that are usually obtained from a MAP-MRF Bayesian approach, classical methods adopted are: Loopy Belief Propagation (LBP) [82], Graph Cuts (GC) [29], Iterated Conditional Modes (ICM) [21], Tree-Reweighted Message Passing (TRW) [94]. A comprehensive analysis and a comparison of such algorithms are presented in [91]. Since usually these methods are adopted in problems which present a global energy function defined for a finite set of variables (sites) which can take discrete values, they are not directly suited for the optimization of the energy function that is derived in this thesis. In fact, such a energy function is defined for a finite set of variables which can take a finite set of values that are sampled from a continuous distribution (each variable takes different values).

It is important to notice that for each point $p_i \in \Lambda_T$ the posterior is calculated on the N_i points $p_i^1, p_i^2, ..., p_i^{N_i}$ sampled with the strategy proposed in the previous section, by accounting the N_j samples of the 4-neighbors $p_j^1, ..., p_j^{N_j}, j : p_j \in \mathcal{N}(p_i)$. Therefore, for each point it is different the number of samples for which the posterior is calculated, and each point corresponds to potentially different distances. The maximization of the posterior expressed in Equation (5.16) can be performed by means of LBP, since it is possible to define the messages of the LBP algorithm also in this situation, in which each site presents a finite set of different labels. As for the classical situation (LP2 problem), it is not given a formal proof of the convergence and of the effectiveness of the application of the LBP algorithm, however the experimental results presented in the next section constitute an evidence of the suitability of LBP for the maximization of the posterior expressed in Equation (5.16). In particular, the messages that the points $p_j \in \mathcal{N}(p_i)$ send to the points $p_i \in \Lambda_Z$ for the distance $z_i^{n_i}$ at the $(t+1)^{th}$ iteration are defined similarly to the classical LBP messages as:

$$m^{t+1}_{p_j \to p_i}(z_i^{n_i}) = \sum_{n_j=1}^{N_j} P^{(data)}(z_j^{n_j}) P^{(smooth)}(z_i^{n_i}, z_j^{n_j}, Th) \prod_{l:p_l \in \mathcal{N}(p_j)-\{p_i\}} m^t_{p_l \to p_j}(z_j^{n_j}) \tag{5.17}$$

All the messages are initialized at 0 before the first iteration: $m^0_{p_j \to p_i}(z_i^{n_i}) = 1, \forall p_j \in \Lambda_Z, \forall p_j \in \mathcal{N}_1(p_i), \forall n_i \in [1, ..., N_i]$. The adopted message updating rule is synchronous. Let us remember that the goal of LBP is the marginalization of the a-posteriori probability for the depth measurements $z_i^1, ..., z_i^{N_i}$ at each site $p_i \in \Lambda_Z$, and then the maximization becomes a winner-takes-all algorithm on the marginalized a-posteriori probability $\hat{P}_i(z_i^{n_i})$ [24]. The final expression of the marginal a-posteriori probability $\hat{P}_i(z_i^{n_i})$ is obtained as:

$$\hat{P}_i(z_i^{n_i}) = \frac{1}{\mathcal{Z}} P^{(data)}(z_i^{n_i}) \prod_{j:p_j \in \mathcal{N}(p_i)} m^{\infty}_{p_j \to p_i}(z_i^{n_i}) \tag{5.18}$$

where $m^{\infty}_{p_j \to p_i}(z_i^{n_i})$ is the value of the message at the last considered iteration of LBP.

5.8 Experimental Results

In order to asses the quality of the proposed ToF fusion framework for data acquired by a ToF camera and a stereo pair, we considered an acquisition setup made by a ToF camera and two standard BASLER scA1000$^{\text{TM}}$RGB cameras $\{L, R\}$, with $4.5mm$ optics, that acquire RGB images $\{I_L, I_R\}$ with resolution 1032×778. The stereo pair S has a baseline of $170[mm]$. The MESA SR4000 ToF camera $\{T\}$, with a $10mm$ optics and horizontal field of view of 43.6^o acquires a 16-bit depth image D_T, with values in $[0, 5m]$, a 16-bit amplitude image A_T, and a confidence map C_T with integer values in $[0, 8]$. Data $\{A_T, D_T, C_T\}$ are framed with resolution 176×144. The ToF camera is positioned in between L and R. The three cameras are synchronized via hardware by a synchronization circuit [2] and the overall acquisition frame rate is $15[fps]$ (its limit is $31[fps]$). The fusion framework takes in input:

- the high resolution ($[1032 \times 778]$) color image I_L acquired by the camera L
- the high resolution ($[1032 \times 778]$) color image I_R acquired by the camera R
- the low resolution ($[176 \times 144]$) depth-map D_T acquired by the ToF camera T
- the low resolution ($[176 \times 144]$) amplitude image A_T acquired by the ToF camera T

The output of the fusion algorithm is a depth-map $\hat{Z}$ with lateral resolution $[352 \times 288]$ (which can go up to $[1056 \times 864]$) of the framed scene, from the point of view of the ToF camera (defined on the lattice Λ_Z), characterized also by high distance measurement accuracy. In order to test the accuracy of the fusion algorithm results, we compared the quality of the estimated depth-map $\hat{Z}$ with respect to a ground truth acquired with a *space-time* stereo vision system [49, 101] for different scenes. For each scene, a set of 600 frames with 600 different projected patterns have been acquired. The ground truth depth-map has been estimated integrating all the 600 images with the 600 patterns, by also applying a sub-pixel refinement and a *left-right check*. The precision of the depth-maps obtained with such a system is of about $1 - 2[mm]$. In particular, the five scenes of Figure 5.11 have been considered for the analysis of the results presented in this chapter.

One of the major contributions of this fusion method is the likelihood model for ToF cameras measurements. The ToF likelihood $P(I_T|Z)$ accounts for the matricial nature of ToF cameras and for depth discontinuities and near IR reflectivity of the scene. In order to validate the correctness of the proposed ToF camera model, we show some examples of ToF likelihood, stereo likelihood and of their multiplication (the so-called data term or joint likelihood). In particular, we show how the proposed model of the ToF likelihood, when combined with the one of the stereo likelihood, allows to improve the accuracy of depth measurements far from depth discontinuities and allows to correct erroneous measurements of the ToF camera in presence of depth discontinuities. Let us firstly consider the case of a pixel p_i far from depth discontinuities in the first scene of Figure 5.11. For such a point, showed in picture 5.12, both the ToF likelihood and the stereo likelihood are calculated with the proposed methods, and such likelihood are reported in the first row of Figure 5.13 and in the second row of Figure 5.13. The data term probability is then calculated from their multiplication and reported in the third

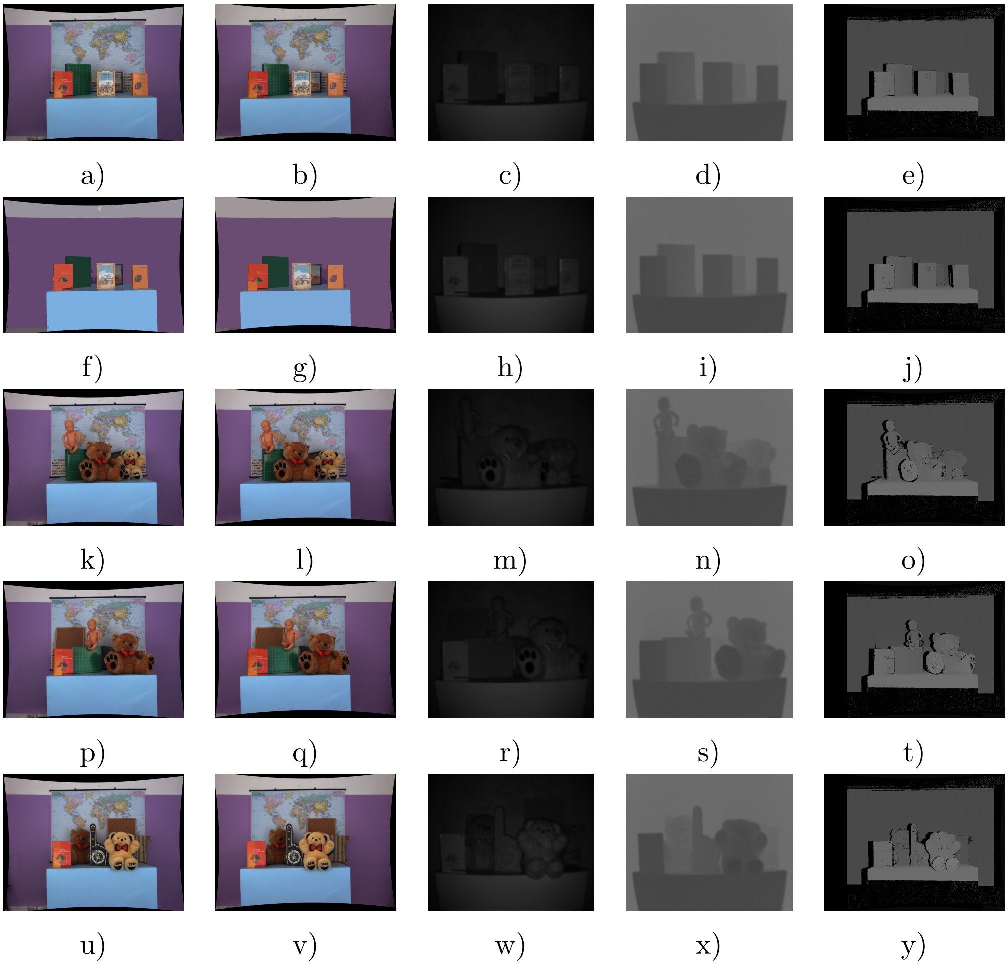


Figure 5.11: Five considered scenes (in the five rows). In the first column I_L (resolution [1032 × 778]), in the second I_R (resolution [1032 × 778]), in the third A_T (resolution [176×144]), in the fourth D_T (resolution [176×144]) and in the fifth scene the ground truth depth-map Z (resolution [1032 × 778]). For the analysis of the results, only the central portion of the acquired scene is considered.

row of Figure 5.13. It is worth to notice how the accuracy of the combined probability is better than the one of the ToF likelihood. In particular, the maximum of the ToF likelihood is relative to the distance 1564$[mm]$, the maximum of the stereo likelihood is relative to the distance 1578$[mm]$, the maximum of the data term probability is relative to the distance 1578$[mm]$, and the ground truth distance is 1580.2$[mm]$.

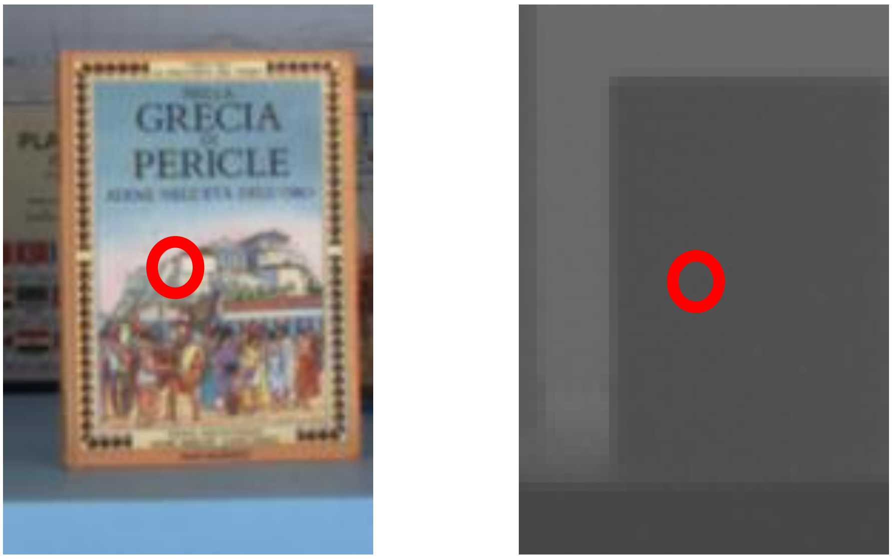


Figure 5.12: Point far from scene distance discontinuities considered in the analysis of the ToF model.

Moreover, let us consider the case of a pixel p_i near depth discontinuities in the first scene of Figure 5.11. For such a point, showed in picture 5.14, both the ToF likelihood and the stereo likelihood are calculated with the proposed methods, and such likelihoods are reported in the first row of Figure 5.15 and in the second row of Figure 5.15. The data term probability is then calculated from their multiplication and reported in the third row of Figure 5.15. In this case, the ground truth distance of the point is 1576$[mm]$. The points is near a distance discontinuity that is characterized by one surface at 1584$[mm]$ (that is the one on which the point actually lies) and one surface at 2079$[mm]$. The point measured by the ToF is 1789$[mm]$. Such an erroneous measurement is due to the effects explained in Section 5.3. The distance measurement performed by the ToF relatively to this point is therefore very imprecise (characterized by an error of 205$[mm]$). Since this point is relative to a very textured surface, the

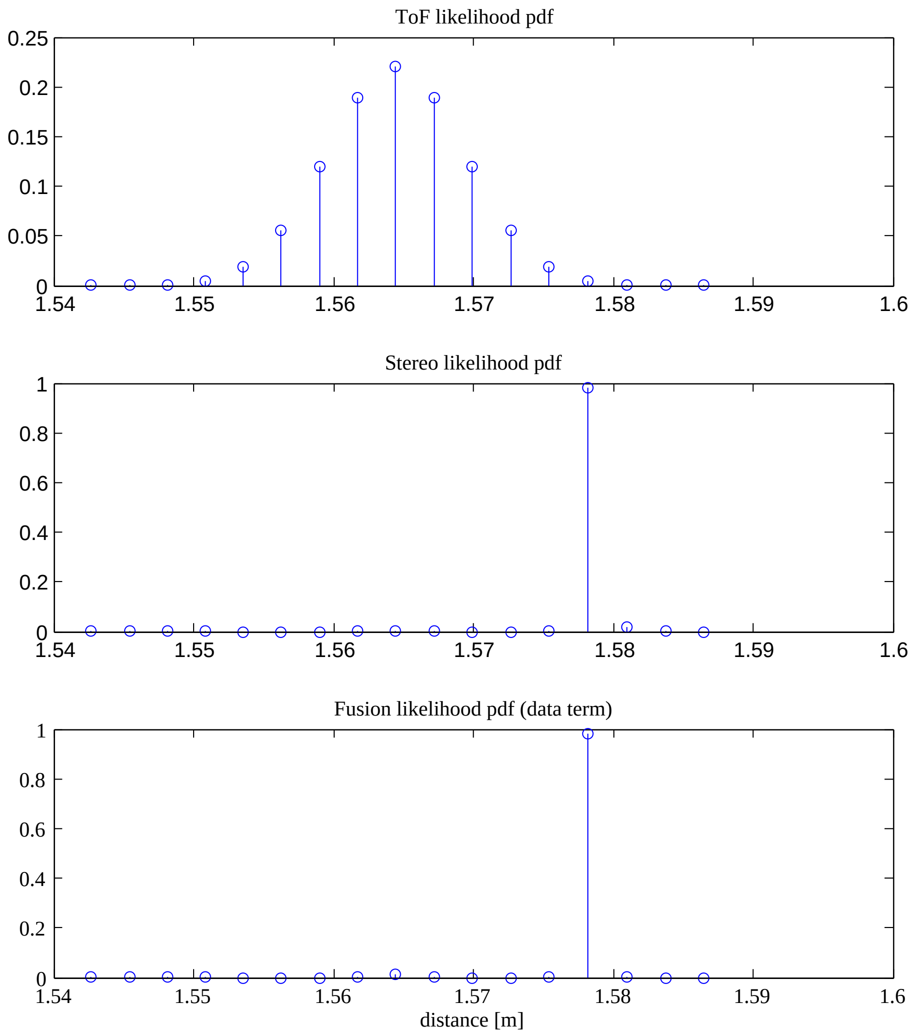


Figure 5.13: ToF likelihood (first row), stereo likelihood (second row) and joint likelihood (third row) relative to a point far from scene distance discontinuities

stereo is very precise in the measure of its distance. In fact, the maximum of the stereo likelihood is at $1576[mm]$. The maximum of the multiplication of the stereo likelihood and of the ToF likelihood is at $1576[mm]$ as well. Therefore, also in this case it is possible to notice how the likelihood of the ToF allows for the compensation of the ToF measurement error by the taking into account also for the stereo likelihood. At this point it is clear that the fusion algorithm allows to improve the quality of the ToF measurements, but it is not clear yet if it is able to improve also the quality of the stereo distance measurements. In order to show how the fusion algorithm allows the improvement of the stereo measurement accuracy, let us consider the point presented in Figure 5.16. The specific point is characterized by the lack of texture in the color images I_L and I_R, therefore the depth measurements performed by the stereo pair result to be very imprecise. The stereo likelihood, shown in the second row of Figure 5.17, does not present a any peak. The ToF measurements however are not affected by the absence of texture. The maximum of ToF likelihood, shown in the first row of Figure 5.17 is relative to a distance of $1410.2[mm]$, that is close to the actual distance of $1411[mm]$. Since in the stereo likelihood does not present a peak, the ToF likelihood shape dominates the data term, and the accuracy of the fusion is the same of the one of the ToF, since the maximum of the joint likelihood is relative to a distance of $1402[mm]$ as well. From this case it is possible to notice how the data fusion algorithm allows to obtain better results than the ones obtained with the application of a stereo vision algorithm alone. Therefore it is clear how the proposed framework adopted for the

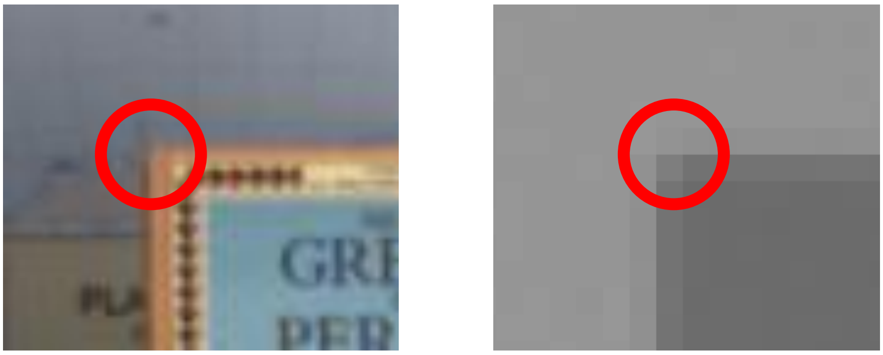


Figure 5.14: Point near scene distance discontinuities considered in the analysis of the ToF model.

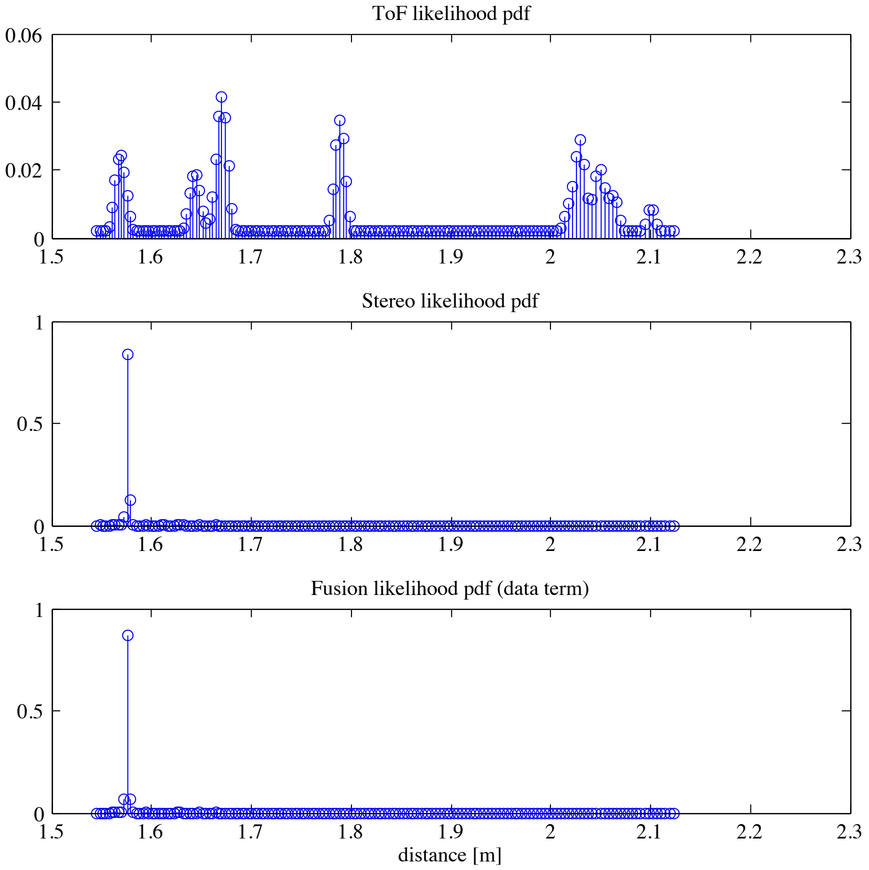


Figure 5.15: ToF likelihood (first row), stereo likelihood (second row) and joint likelihood (third row) relative to a point near scene distance discontinuities.

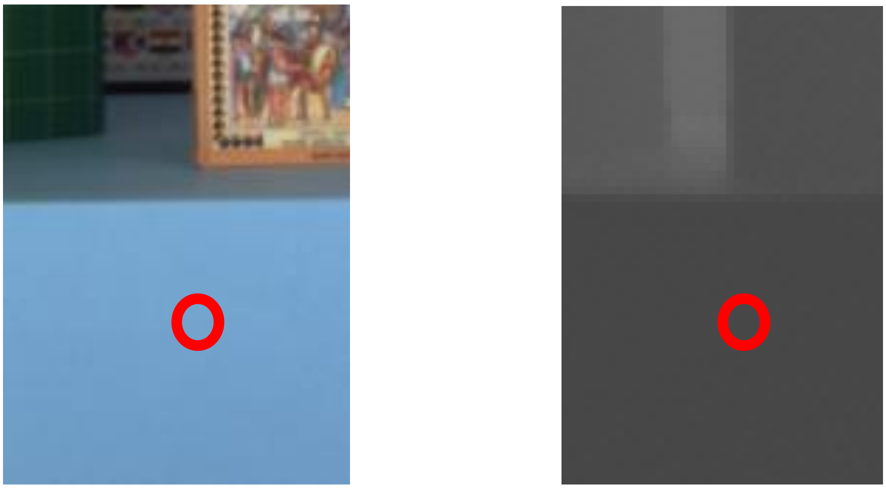


Figure 5.16: Point in a texture-less area.

calculation of the joint likelihood by the exploitation of the ToF model works robustly with respect to scene depth discontinuities and textured and textureless surfaces.

In order to see the effectiveness of the computed joint likelihood it is possible to consider for the first scene in Figure 5.11 the maximum of the data term for each single point as the estimated distance, disregarding the MRF assumption for $\mathcal{Z}$ (in this case the depth measurement of the various pixels are assumed independent). This is equivalent to adopting a Maximum-Likelihood (ML) approach. The results of such an approach are not obtained by applying a global optimization algorithm such as LBP, but just by picking the distance for each pixel that maximizes the joint likelihood for the pixel itself. In Figure 5.18 the distance error obtained with the ML approach for the first scene of Figure 5.11 is reported. From the error map reported in Figure 5.18 it is immediate to notice how the major errors are relative to the texture-less slanted surface of the table. In fact, both the measurement of the ToF and the stereo on such a surface are not very precise, since the surface is slanted (this affects the quality of the ToF measurements) and texture-less (this affects the quality of the stereo measurements). However, it is possible to notice how there is an overall improvement after the application of the fusion algorithm in the distance measurement accuracy of the ToF and the stereo. In fact, the error of the ToF measurements is of $22.2[mm]$, the error of the stereo measurements is of $30[mm]$ and the error after the application of the fusion algorithm is of $20[mm]$. It is important to notice that the error of the stereo measurement is not comparable

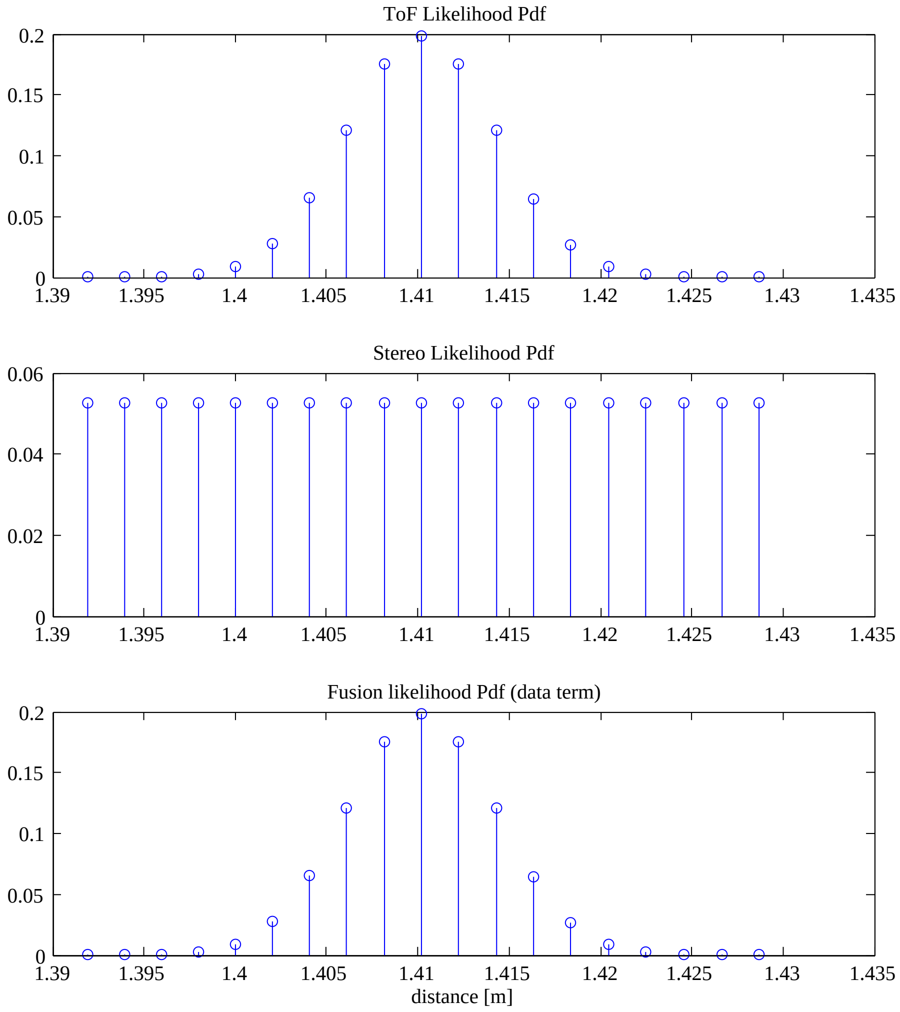


Figure 5.17: ToF likelihood (first row), stereo likelihood (second row) and joint likelihood (third row) relative to a point in a texture-less area.

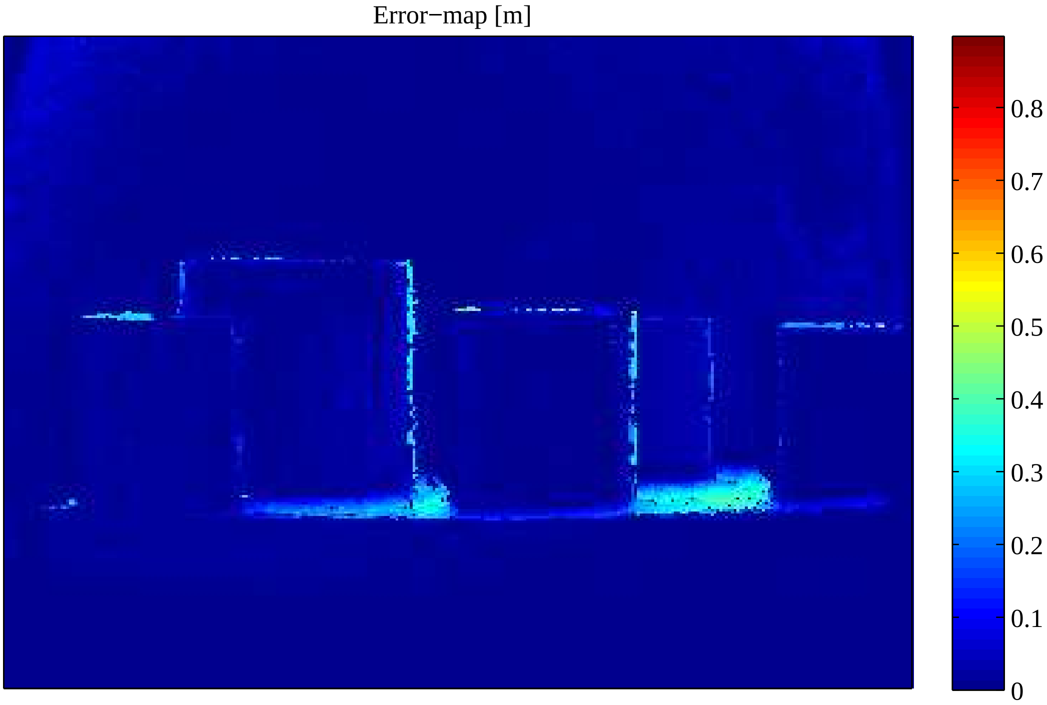


Figure 5.18: Error-map after the application of the ML optimization, without accounting for the MRF assumption. The map and the color-bar are both expressed in $[m]$.

to the error of a classical stereo vision algorithm, because it is affected by the great improvement introduced by the concept of *useful interval*. Since the error in Figure 5.18 is dominated by the table surface error, it is possible to perform a further analysis of the improvement in the distance measurement accuracy from the ToF measurements only to the results of the fusion algorithm (without the application of LBP optimization), by considering the accuracy only of the scene region reported in Figure 5.19. With respect to such a region, it is possible to observe that the average error of the ToF measurements is 18[mm], while the average error of the fused data is 15[mm]. The accuracy of the ToF measurements is increased by 17% after the application of the fusion algorithm. If the MRF hypothesis is considered and the LBP optimization algorithm is adopted, it is interesting to notice how the accuracy of the distance estimation increases. This shows the effectiveness of the proposed extension of the LBP algorithm. In order to provide a more complete set of results, we also tested the proposed data fusion method on the five scenes reported in Figure 5.11. For each scene, the following quantities are reported in Table 5.1:

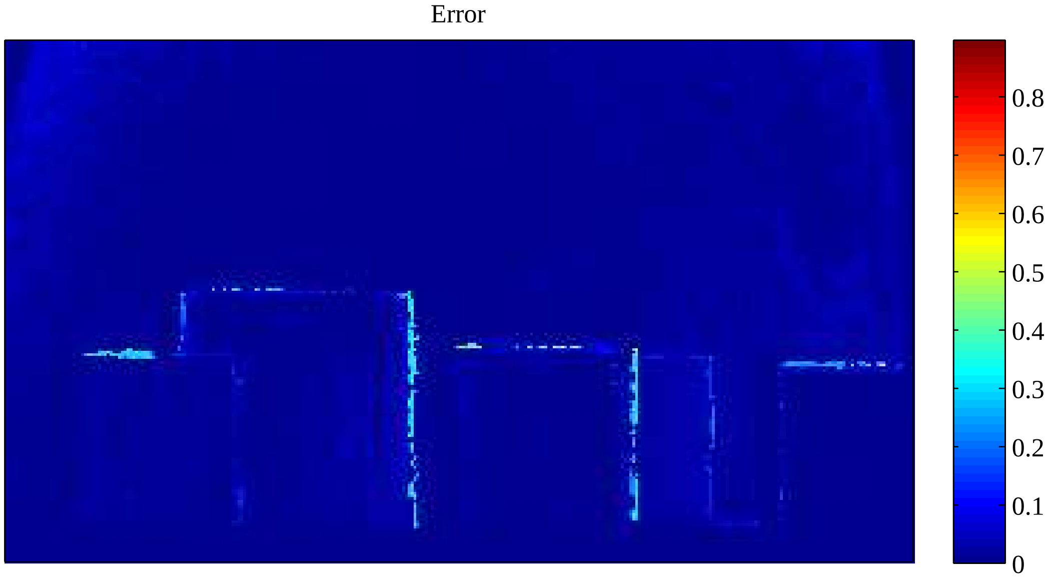


Figure 5.19: Particular of the error map after the ML optimization. The map and the color bar are expressed in [m].

- the ToF accuracy (defined as in Chapter 4)

Scene	ToF Accuracy	Stereo Accuracy	ML Accuracy	MAP Accuracy
1	22	30	20	17
2	24	35	20	18
3	27	30	25	23
4	28	29	22	20
5	27	31	25	24

Table 5.1: Accuracy of depth information acquired with ToF, stereo, ML fusion and MAP fusion approach. Accuracy of the ML fusion approach is always better than the one of ToF and stereo measurements. Accuracy of the MAP fusion approach is always better than the one of the ML fusion approach. Stereo accuracy accounts for the *useful interval* concept.

- the stereo accuracy (defined as in Chapter 4)
- the accuracy of the obtained depth-map by applying the MAP criterion, with the proposed LBP optimization algorithm (defined as in Chapter 4)

In Figure 5.20 the estimated depth-maps (MAP-approach) for the five scenes of Figure 5.11 are presented.

In Figure 5.21 it is reported the MAP estimation error as a function of LBP iterations for the fifthscene.

Figure 5.22 reports a particular on the differences in the estimate of the depth-map for the third scene before and after the application of LBP, *i.e.*, the ML and the MAP estimates.

The accuracy of the proposed method is also comparable or better than the accuracy of ToF and stereo systems and the depth resolution is less than $1[mm]$ (it can be tuned by a different sampling inside the *useful interval.*)

Unfortunately the comparison with other methods for ToF and stereo data fusion is not trivial due to the unavailability of common datasets or other method's code. Given the enormous amount of details that characterize each approaches, an exhaustive implementation is not feasible with reasonable effort. Therefore it has been chosen not to present any quantitative comparison. With respect to some methods, it is possible to perform a qualitative analysis, which is presented hereafter. In particular, with respect to [42], the proposed method allows to provide an high resolution depth-map instead of the one at the low resolution of the ToF camera. Moreover, the global optimization step allows to provide a more robust estimate of the scene depth distribution w.r.t. the

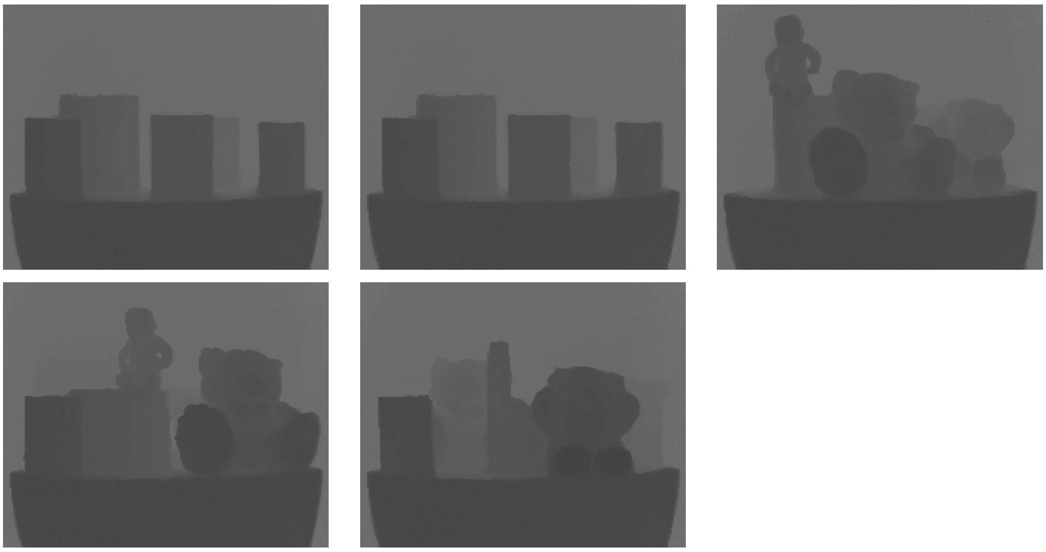


Figure 5.20: MAP estimates of the depth-maps for the five scenes of Figure 5.11. In order not to create artifacts due to interpolation, the estimated depth-maps are not compensated for camera distortion. It is possible to perform the undistortion artifacts-free directly in the three-dimensional space.

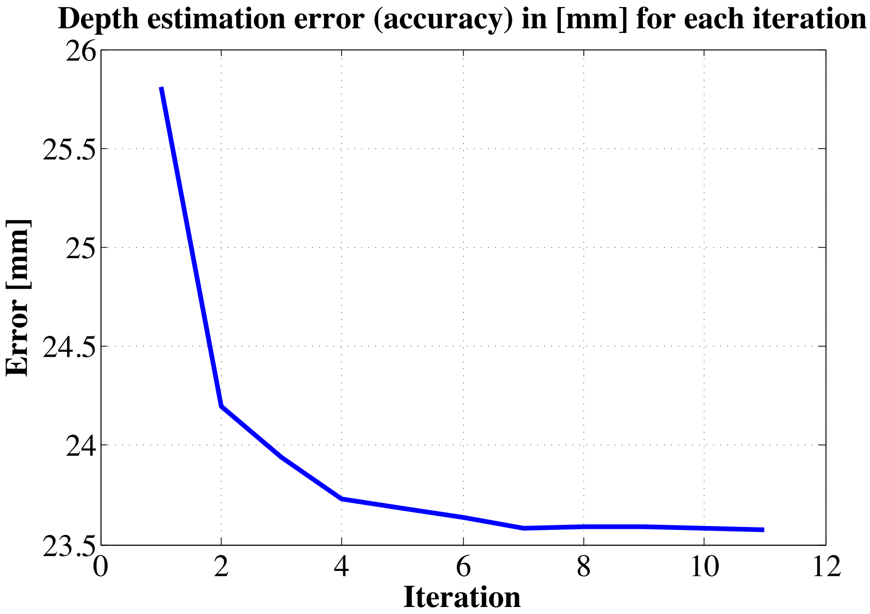


Figure 5.21: MAP estimation error as a function of LBP iterations for the fifth scene.

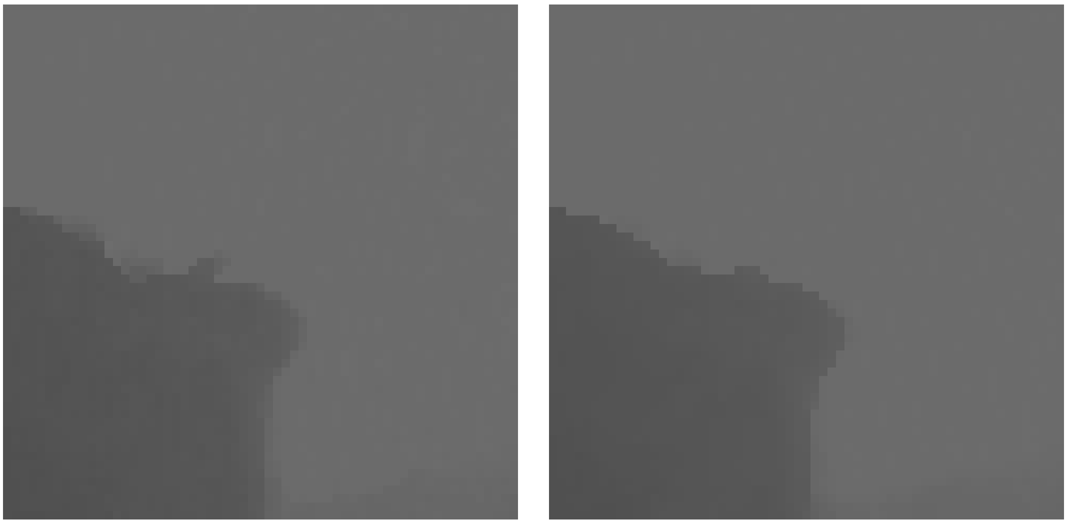


Figure 5.22: Particular of the estimated depth-map with ML (left) and with MAP (right) approaches. The application of LBP improves the quality of the estimated depth-map.

methods of [42, 98, 99]. The most interesting comparison of the proposed method is the one with [104], since the two frameworks are very similar. It is interesting to notice how the quality of the two methods is comparable, even though our method is more accurate than [104]. This accuracy improvement is due to the rigorous ToF likelihood derivation, obtained by the analysis of ToF cameras and to the *useful interval.* It is moreover necessary to remember that the *useful interval* restriction allows our method to perform 7% of the operations of the method proposed in [104].

6

Scene segmentation from 3D and color data

6.1 Introduction

Scene segmentation is the well-known problem of identifying the different elements of a scene. Images are the most common way of representing scenes, therefore it is not surprising that scene segmentation by way of images has attracted a lot of attention. Unfortunately scene segmentation by images is an ill-posed problem, and, despite a huge amount of research, it is still a very challenging task. Many segmentation techniques based on different insights have been developed, such as methods based on graph theory [52], methods based on clustering algorithms, (*e.g.* [36] and [87]), and also other methods based on region merging, level sets, watershed transforms and many other techniques [90]. The main drawback of image segmentation, independently from the deployed technique, is that the information carried by a single image may not suffice to completely understand the scene structure (consider for instance the simple case of an object and a background of the same color). As shown in previous chapters, current technology allows to acquire scene descriptions beyond simple images. Besides stereo vision systems and ToF cameras, structured-light cameras (*e.g.*, Microsoft Kinect [9]) have reached the market and are gaining popularity. Unstructured scene reconstruction tools like Microsoft Photosynth [10] can also provide the geometrical representation of a scene from a collection of pictures taken from random positions. The fusion of depth information acquired by any of these tools together with the color information coming from a standard color camera allows to obtain scene descriptions accounting for both geometry and color, *i.e.*, representations where each sample has both geometry and color information associated to it. In this context, scene segmentation can be approached within a sensor fusion framework by algorithms exploiting both clues together and not just color as in standard segmentation algorithms. Within this perspective the segmentation problem can be formulated as the search for effective ways of meaningfully partitioning a set of samples featuring color and geometry information.

While the literature about scene segmentation based on color information is extremely vast, the number of works addressing scene segmentation by way of color and geometry information is still rather limited. A first possible solution is to perform two independent segmentations, one on the color image and one on the depth data, and then join the two results, as proposed in [34]. Many approaches, like [60] and [74], consider the special case of the recognition of the foreground from the background rather than the general scene segmentation case. In [96] two likelihood functions, one built on the basis of depth information and the other on the basis of color data, are combined together in order to assign samples to the background or to the foreground. Two different approaches for the segmentation of binocular stereo video sequences are presented in [70]: one, based on Layered Dynamic Programming, explicitly extracts depth information while the other one, based on Layered Graph Cuts, uses stereo correspondences without explicitly computing depth. Some other recent works try to jointly solve the segmentation and stereo disparity estimation problems. Ladicky et al. [72] exploit a probabilistic framework based on Conditional Random Fields. This approach uses some heuristics about the scene structure that limit it to a particular scene setting (*i.e.*, urban streets). A more general approach, also based on a probabilistic framework has been presented in [27].

Clustering techniques has been widely used in image segmentation and are well-suited to be extended in order to include different spatial and color features as shown in [79]. They can be exploited for joint depth and color segmentation by adding also the depth component to the vectors that are then clustered. Bleiweiss et Werman [26] follow this approach and apply *mean shift* clustering to vectors containing both the color and depth information. In [95] superparamagnetic clustering and channel representations are instead exploited to segment plant scenes from the color and depth data acquired by a Microsoft Kinect camera.

This chapter proposes a novel general scene segmentation scheme based on normalized cuts spectral clustering [87], which exploits the fusion of geometry and color information in a parameterless framework. It is proposed a completely general approach that can be applied in a fully automated way (*i.e.* it does not require any supervision for the choice of the balancing parameter between depth and color) regardless of the acquisition device and data type.

The chapter is organized as follows: Section 6.2 formalizes the adopted scene representation fusing both color and geometry. Section 6.3 introduces the proposed scene segmentation algorithm based on the normalized cuts spectral clustering algorithm. In Section 6.4 an algorithm for the automatic balancing of the weight between geometry and color is proposed. It is based on a novel unsupervised metric for scene segmentation quality assessment. Section 6.5 proposes an extension of the segmentation algorithm tailored to the important case of stereoscopic data that besides geometry exploits the color of both images of a stereo pair. Section 6.6 reports the experimental results and demonstrates how the joint exploitation of geometry and color within the proposed method outperforms segmentation algorithms based on either geometry or color information only, or on the joint exploitation of the two clues. In Section 6.7 the results of the segmentation of the same scene acquired with different depth imaging techniques are presented and the performance of the different acquisition systems for segmentation purposes are discussed.

6.2 Joint representation of geometry and color information

Figure 6.1 shows an overview of the proposed scene segmentation algorithm. The procedure can be subdivided into two main stages. In the first stage, a unified 6-dimensional representation of the scene points is built in order to fuse geometry and color information in a fully automatic way. In the second stage the obtained point set is segmented by means of spectral clustering.

This section addresses the construction of the unified representation for the joint exploitation of geometry and color information. The description assumes the availability of a generic scene $\mathcal{S}$ described by a set of N points $p_i, i = 1, ..., N$ featuring both

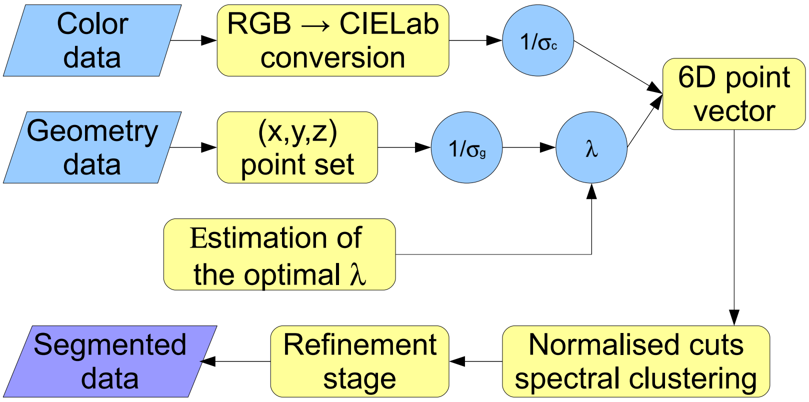


Figure 6.1: Architecture of the proposed segmentation scheme

geometry and color information. Let us stress that for our purposes, the specific characteristics of the used 3D acquisition system are irrelevant and the acquired scene can be represented both by an image with the corresponding depth map or by a colored sparse point-cloud independently of the acquisition system. Such independence from the acquisition equipment is of major practical relevance since it allows to apply the proposed segmentation method with total generality to any type of color and geometry data describing a scene.

Color data require a 3D vector, in order to account for the R, G and B color components and another 3D vector is required for geometry information in order to describe the x, y and z coordinates of a point with respect to a given reference system (such a reference system can be obtained from the calibration data and the depth-maps produced by many acquisition systems). First of all geometry and color information need to be unified in a meaningful way. We choose to represent the color values in a perceptually uniform space in order to give a perceptual significance to the Euclidean distance between colors. This helps keeping consistent with the perceived color difference the distances used in the clustering process of Section 6.3. Note also that a uniform color space ensures that the distances in each of the 3 color components are comparable, thus simplifying the clustering of the 3D vector associated to color information. The CIELab space was selected for color representation, *i.e.*, the color information of each scene point $p_i, i = 1, ..., N \in \mathcal{S}$, is the 3D vector:

$$\mathbf{p_i^c} = \begin{bmatrix} L(p_i) \\ a(p_i) \\ b(p_i) \end{bmatrix} \quad , i = 1, ..., N \tag{6.1}$$

Geometry can be simply represented by the 3D coordinates $x(p_i)$, $y(p_i)$, and $z(p_i)$ of each point $p_i \in \mathcal{S}$, *i.e.* as:

$$\mathbf{p_i^g} = \begin{bmatrix} x(p_i) \\ y(p_i) \\ z(p_i) \end{bmatrix} \quad , i = 1, ..., N \tag{6.2}$$

An ideal scene segmentation algorithm should be insensitive to the relative scaling of the point-cloud geometry since not all the scene acquisition systems are able to provide geometrical descriptions with respect to an absolute scale system (*e.g.* meters). For instance, tools like Photosynth [10] are only able to reconstruct the scene geometry up to an arbitrary scale factor. Therefore, in order to be independent with respect to scaling, all the components of $\mathbf{p_i^g}, i = 1, ..., N$ are normalized w.r.t. the average σ_g of the standard deviations of the point coordinates. To be more precise, let σ_x, σ_y and σ_z be the standard deviations of sets $x(p_i)$, $y(p_i)$ and $z(p_i), i = 1, ..., N$ respectively. The average standard deviation is then defined as $\sigma_g = (\sigma_x + \sigma_y + \sigma_z)/3$ and the adopted geometry representation is vector:

$$\begin{bmatrix} \bar{x}(p_i) \\ \bar{y}(p_i) \\ \bar{z}(p_i) \end{bmatrix} = \frac{3}{\sigma_x + \sigma_y + \sigma_z} \begin{bmatrix} x(p_i) \\ y(p_i) \\ z(p_i) \end{bmatrix} = \frac{1}{\sigma_g} \begin{bmatrix} x(p_i) \\ y(p_i) \\ z(p_i) \end{bmatrix} \tag{6.3}$$

It is worth to notice that since the proposed segmentation algorithm is based on relative points distances and the overall distances are normalized, segmentation based on (6.3) besides scaling will also be insensitive to the choice of the reference frame. Furthermore by using the coordinates of the point in the 3D space it is ensured that all the three spatial dimensions refer to the same space and that they are consistent, differently from other approaches like [26] where the 2D coordinates in image space are used together with depth data, which lies in a different space.

In order to balance the relevance of the two kinds of information (color and geometry) in the merging process, color information vectors $\mathbf{p}_i^c, i = 1, ..., N$ are normalized as well by the average σ_c of the standard deviations σ_L, σ_a and σ_b of their L, the a and the b components respectively. The final color representation therefore is:

$$\begin{bmatrix} \bar{L}(p_i) \\ \bar{a}(p_i) \\ \bar{b}(p_i) \end{bmatrix} = \frac{3}{\sigma_L + \sigma_a + \sigma_b} \begin{bmatrix} L(p_i) \\ a(p_i) \\ b(p_i) \end{bmatrix} = \frac{1}{\sigma_c} \begin{bmatrix} L(p_i) \\ a(p_i) \\ b(p_i) \end{bmatrix} \tag{6.4}$$

Given the above normalized geometry and color information vectors, each scene point $p_i^f, i = 1, ..., N$ is represented as:

$$\mathbf{p}_i^f = \begin{bmatrix} \bar{L}(p_i) \\ \bar{a}(p_i) \\ \bar{b}(p_i) \\ \lambda\bar{x}(p_i) \\ \lambda\bar{y}(p_i) \\ \lambda\bar{z}(p_i) \end{bmatrix}, i = 1, ..., N \tag{6.5}$$

where λ is a parameter balancing the contribution of color and geometry. High values of λ increase the relevance of geometry, while low values of λ increase the relevance of color information. Figure 6.2 shows an example of the relevance of λ in the segmentation of the *plant* scene, which is a 3D model obtained by Microsoft Photosynth. For low values of λ (*e.g.*, $\lambda = 0.001$) the segmentation is dominated by the color clue, thus leading to some artifacts due to the noise on the color data. For higher value of λ (*e.g.*, $\lambda = 5$), the segmentation is dominated by the geometry clue, and the entire plant is segmented into three parts that do not take in account color, denying as well a meaningful segmentation. For intermediate values of λ (*e.g.*, in this case $\lambda = 1$), geometry and color information in this case are well balanced providing correct segmentation results by the proposed method. Note that the value of λ leading to the best segmentation results depends on the specific scene data.

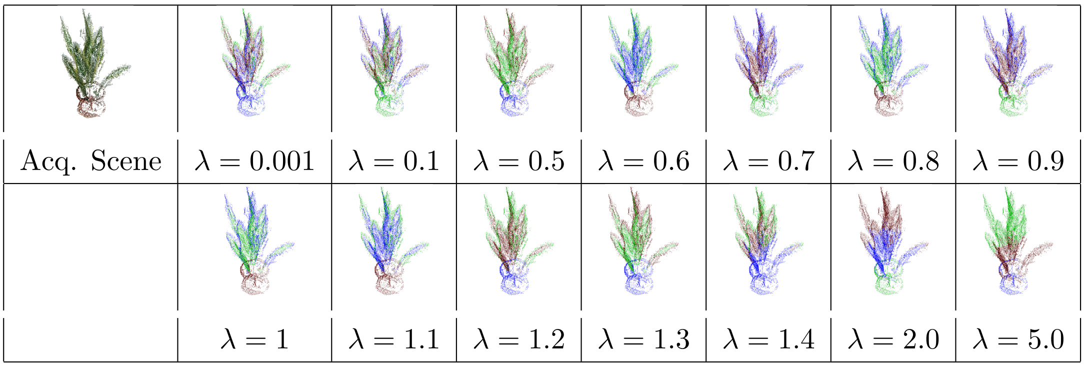


Figure 6.2: Different segmentation results on the *plant* scene for different values of λ. *(Best viewed in colors)*

6.3 Segmentation by means of spectral clustering and Ny ström method

The representation of a scene introduced in the previous section is characterized by a set P_c formed by the 6D vectors $\mathbf{p}_i^f, i = 1, ..., N$ which represents in a intuitive and consistent way the geometry and color information of the scene points $p_i, i = 1, ..., N$. Vectors $\mathbf{p}_i^f$ are well suited for clustering. Central grouping algorithms, such as k-means and mean-shift clustering, are fast and effective, but have the main drawback of assuming specific distributions of the points in P_c. Since this assumption is not generally verified in the considered application, this family of methods applied to the set P_c gives poor results. Figure 6.3 shows an example of the results of k-means clustering and of mean-shift clustering on set P_c of points relative to the *baby* scene. The methods based on pairwise affinity measures computed between all the possible couples of points in P_c operate somehow within a philosophy opposite to that of central grouping. They are more flexible, because they do not assume a specific model for the distribution of the points, and consequently their results in practical segmentation situations are more accurate and robust. The main drawback of the pairwise affinity methods is that they need to compare all the possible pairs of points in P_c. Computing and storing all the possible affinities forces a tremendous amount of processing, very expensive in terms of both CPU and memory resources. Normalized cuts spectral clustering presented in [87] is an outstanding example within this family. In this method a graph is firstly built

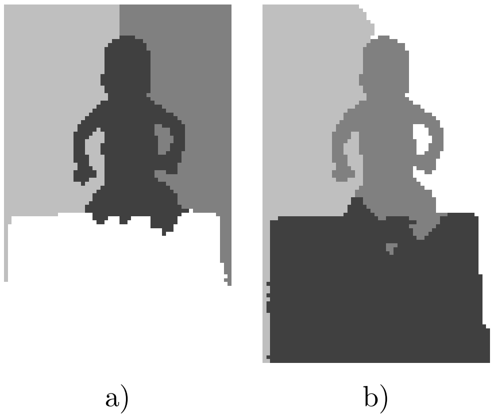


Figure 6.3: Segmentation of the *baby* scene applying a) k-means clustering and b) mean-shift clustering

from all the points (vertices) and their pairs (edges), and then partitioned according to spectral graph theory criteria. Normalized cuts is the minimization criterion adopted for the graph cut in this case in order to account both for the similarity between the pixels inside the same segment and the dissimilarity between the pixels in different segments. The minimization by the normalized cut criterion can be regarded and solved as a generalized eigenvalue problem. A variety of methods have been proposed for the efficient approximation of the graph associated to the set of points in order to overcome the computational and memory burden. A possible solution is imposing that not all the points are connected, but that the non negligible connections only concern small sets of points. This assumption practically leads to oversegmentation, and implicitly imposes some models to the point distributions. In the method based on the integral eigenvalue problem proposed in [53] the set of points is firstly randomly subsampled (a set of n points is randomly extracted from the whole set of N points); this subset of n points is then partitioned by the method proposed in [87], and the solution is propagated to the whole N points set by a specific technique called Nyström method. As shown in [53], the results of this method are comparable to the ones of the normalized spectral clustering algorithm, but at computation and memory costs comparable with those of the central grouping algorithm. For this reason the Nyström method approach to the normalized cut spectral clustering (briefly denoted with NNCSC) was selected for our scene segmentation application. The fact that NNCSC does not assume any model for the distribution of the points in P_c is a rather important feature. In some way, NNCSC provides a nice framework to incorporate the fact that P_c is partitioned into subsets where color and geometry are homogeneous, without imposing an overall model, which for the distributions of the points in P_c would be very hard to derive. For a detailed explanation of normalized cuts spectral clustering, the interested reader is referred to [87], and for Nyström method to [53]. A drawback of normalized cuts, shared with other clustering algorithms like k-means, is that the number of clusters K in which the point-cloud is partitioned needs to be known a priori. This issue can be overcome by the use of an automatic selector of the number of clusters K, such as the one proposed in [85]. The Nyström method approximation leads to a very fast algorithm, hence suitable for real time applications. Within the following experimental validation, it is shown that the clusters found by NNSC applied to P_c represent rather well the different scene regions.

In order to avoid small regions due to noise it is also possible to include an optional refinement stage for samples arranged on regular grids (*i.e.*, when the input data are images and depth maps) where regions with extension smaller than a threshold are removed and their points are assigned to the cluster corresponding to the mode of the points closer to the region [44]. Such a refinement was instead not used in the results of Figure 6.15, 6.16 and 6.17 where the data are not aligned on regular grids.

6.4 Automatic weighting of color and depth information

The optimal value of the λ parameter, *i.e.* the relative weight between depth and color information, depends on the color and geometry properties of the scene and it turns out to be a key issue in the proposed segmentation scheme. Given that a single optimal value of λ does not exist, this section proposes an effective method for the automatic setting of λ, based on an unsupervised metric for segmentation quality assessment. This approach allows to obtain a parameterless segmentation method that does not rely on manual tuning of the weighting coefficient λ.

A number of unsupervised metrics for the evaluation of image segmentation quality have been proposed in the last decades (a comprehensive taxonomy of them is given in [100]). Among the various metrics of the literature, the FRC metric of [84] has proven to be at the same time very reliable and computationally fast. This method, as proposed by the authors, takes as input a color image and a segmentation map and returns as output a measure of the segmentation quality. Our context is slightly different, because our input is threefold, namely a color image I, a depth-map D (with the geometry information) and a segmentation map S (where the image has been divided in a set of K segmented regions $S_i, i = 1, .., K$) and we are forced to introduce a novel segmentation metric that considers together both color and geometry. In the case of unstructured data representations (*i.e.* point clouds), each point has an associated 3-dimensional color vector and I is simply the set of all the color vectors associated to the 3D points. The depth map D is instead replaced by a set of 3-dimensional vectors with the (x, y, z) coordinates. The segmentation map simply associates each point to one of the clusters. Both color and geometry data are firstly normalized as follows:

- The three color channels (red, green and blue) of I, *i.e.*, I_R, I_G and I_B are normalized in order to obtain a color representation $\tilde{I}$ with values in the range

[0, 1].

- The depth map D is also normalized to depth-map $\tilde{D}$ with values in $[0, 1]$. In the case of unstructured data D is also shifted and normalized in order to have all the coordinates in the range $[0, 1]$. More precisely, for unstructured data, the chosen normalization factor is the maximum of the sides of the bounding box including the point cloud. The same normalization factor is used for all the 3 dimensions in order to avoid "stretching" the point cloud.

Following the approach presented in [84], a "good" segmentation should have two fundamental properties, namely:

- inside a single segmented region the image should have uniform properties (*i.e.*, a constant color or some repeating pattern or texture).
- each couple of different segments should have different properties (this ensures that there is no over-segmentation of the image).

In the considered situation the above criteria should be satisfied with respect both to the color image and to the depth-map. Firstly we consider the segmentation map S and the normalized color image $\tilde{I}$: the evaluation of the first property is quite simple for regions of constant color, where it is usually associated to the standard deviation of the data inside the segmented region, but it is quite difficult for heavily textured regions. This issue in [84] and other works on segmentation evaluation is approached by computing various texture or color distribution descriptors. Unfortunately such descriptors are not always reliable. Indeed heavily textured regions with complex color patterns are where both state-of-the-art segmentation techniques and evaluation metrics usually either have major problems or completely fail. Since in our application also depth information is available, we decided to give more importance to the color component of the metric in regions with limited texture and less importance in heavily textured regions where depth data can be more reliable. The idea adopted to obtain this result is to subtract from the standard deviation of the data of a segmented region the standard deviation due to the amount of texture inside the region. More precisely it is assumed that the amount of texture of a segmented region S_i, denoted as $\sigma_t(S_i)$, is proportional to the average local standard deviation of the samples internal to segment S_i, namely:

$$\sigma_t(S_i) = \frac{\sum_{j \in S_i^*} \sigma_w(j)}{|S_i^*|} \tag{6.6}$$

where $\sigma_w(j)$ is the local standard deviation computed on a small window (for the experimental results a 3×3 window has been used) centered on pixel j. S_i^* is the set of the internal pixels of segment S_i, *i.e.*, the ones for which window $w(j)$ lies completely inside the segment. $|S_i^*|$ is instead the cardinality of S_i^*. Note that this reasoning assumes that the scene color information is represented by way of an image. If the scene is represented by a sparse colored point cloud the window can be replaced by the set of the points with distance from j lower than a threshold t. In the case of point clouds this approach is however computationally expensive. It can be made faster by avoiding the subtraction of the texture standard deviation at the price of a loss in the metric performances. A measure of the internal disparity D^i_{intra} of the i^{th} segment S_i can be computed as follows:

$$D_i^{intra} = \max(\sigma(S_i) - \sigma_t(S_i), 0)\frac{|S_i|}{N} \tag{6.7}$$

where $\sigma(S_i)$ is the global standard deviation of the color data inside the segmented region, $|S_i|$ is the cardinality of the points in the i^{th} region S_i and N is the total number of points in P_c. As previously said the idea is to consider the standard deviation due to the clustering accuracy and not to the complexity of the texture pattern inside the segmented region. The average local standard deviation is therefore subtracted to the global standard deviation of the color inside the region (in the case that the local standard deviation is greater than $\sigma(S_i)$, D_i^{intra} will be set to 0). Expression (6.7) reduces the weight of highly texturized regions, which is quite reasonable in light of the fact that for these regions depth data offer more reliable indications. This is particularly true if depth information is computed by stereo vision techniques since their performance, as well known, is more reliable in textured regions. In any case it seems rather reasonable to use depth in heavily texturized regions and color information in regions with uniform or limited texture which are easy to segment by color information and usually correspond to areas where depth is poorly estimated due to the lack of features to be matched. Finally the segments are also weighted on the basis of their size.

The D^{intra} measure for the whole image is computed as the sum of the D_i^{intra} values of each segmented region:

$$D^{intra} = \sum_i D_i^{intra} \tag{6.8}$$

The disparity between the different segmented regions is instead computed as the distances between the centroids of pairs of clusters (note that here a cluster corresponds to a segmented region) as in the FRC metric introduced in [84]:

$$D_{i,j}^{inter} = |E(S_i) - E(S_j)| \tag{6.9}$$

These disparities are then averaged on all the segment pairs:

$$D^{inter} = \frac{\sum_{i,j(i \neq j)} D_{i,j}^{inter}}{K(K-1)} \tag{6.10}$$

and the final metric for color data is computed as the difference between the disparity between different regions and the internal disparity divided by 2, *i.e.*, as:

$$Q^{color}(\tilde{I}, S) = \frac{D^{inter} - D^{intra}}{2} \tag{6.11}$$

The metric for geometry information is computed in the same way but without considering the local standard deviations, namely:

$$\begin{aligned} D_i^{Dintra} &= \sigma^D(S_i)\frac{|S_i|}{N} && (6.12)\\ D^{Dintra} &= \sum_i D_i^{intra} && (6.13)\\ D_{i,j}^{Dinter} &= |E^D(S_i) - E^D(S_j)| && (6.14)\\ D^{Dinter} &= \frac{\sum_{i,j(i \neq j)} D_{i,j}^{Dinter}}{K(K-1)} && (6.15)\\ Q^{depth}(\tilde{D}, S) &= \frac{D^{Dinter} - D^{Dintra}}{2} && (6.16) \end{aligned}$$

where $\sigma^D(S_i)$ is the standard deviation of the geometry values in region S_i and D^{Dinter} is also computed with respect to geometry data. Note how D is a set of scalar values in the case of depth-maps and a set of 3-dimensional vectors in the case of point clouds, *i.e.* in the unstructured data case D has the same structure of color data with x, y and z in place of the three color channels. Finally the combined segmentation quality metric is computed as follows:

$$Q(\tilde{I}, \tilde{D}, S) = Q^{color}(\tilde{I}, S) + n_f * Q^{depth}(\tilde{D}, S) \tag{6.17}$$

$$\text{with} \quad n_f = \begin{cases} 1 & \text{for unstructured data} \\ 3 & \text{for depth-maps} \end{cases} \tag{6.18}$$

In the case of depth-maps depth relevance is multiplied times 3 in order to assign the same total weight to the 3 color channels together and to the depth data. In the unstructured data case both representations have 3 components and the multiplication by 3 is not needed.

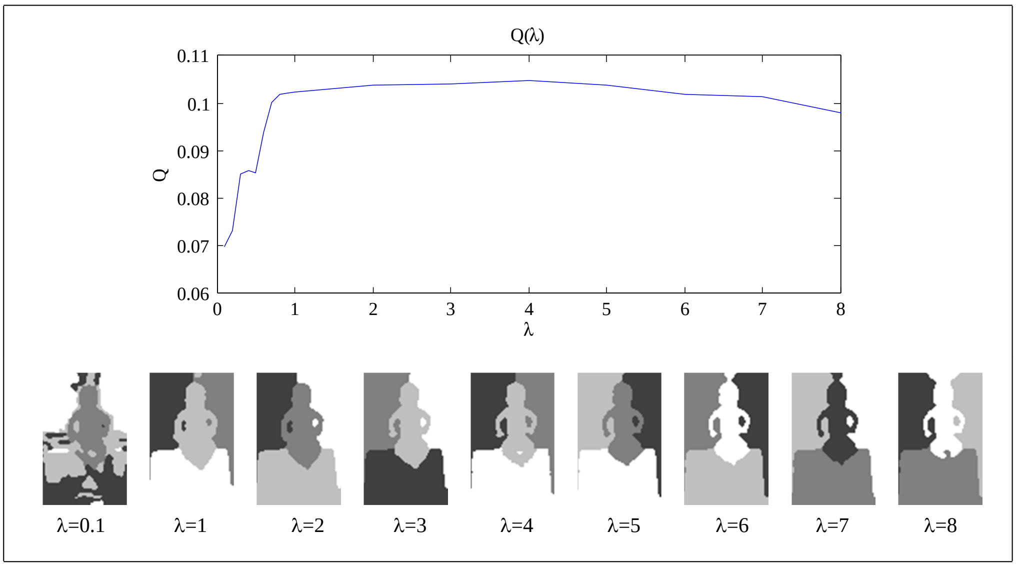


Figure 6.4: Values of Q versus λ and segmentation of the *baby* scene for different values of λ

The optimal λ can be automatically selected as the value that maximizes the $Q(\tilde{I}, \tilde{D}, S)$ value in (6.17). Different values of λ correspond to different segmentation maps S that in turn correspond to different values of $Q(\tilde{I}, \tilde{D}, S)$. The value of λ that maximizes (6.17) is the value that provides the best segmentation with respect to the Q metric. This approach was experimentally found to be very effective, indeed in all the experimental examples it always gave the value of λ providing the best segmentation. An example of this fact is reported in Figure 6.4 where the maximum of Q (obtained for $\lambda = 4$) corresponds to the best segmentation. Indeed only for $\lambda = 4$ even the part of the box between the legs of the baby is correctly associated to the box segment. The plot of Q versus λ clearly shows how the correspondence between the values of λ and the changes in segmentation quality are well reflected by changes of the $Q(\tilde{I}, \tilde{D}, S)$ value. Figure 6.5 shows the behaviour of metric Q versus λ on a different scene, while Figure 6.6 refers to the computation of the metric on a point cloud representation instead of a color image and a depth-map as in the other two cases. It is worth noting that, al-

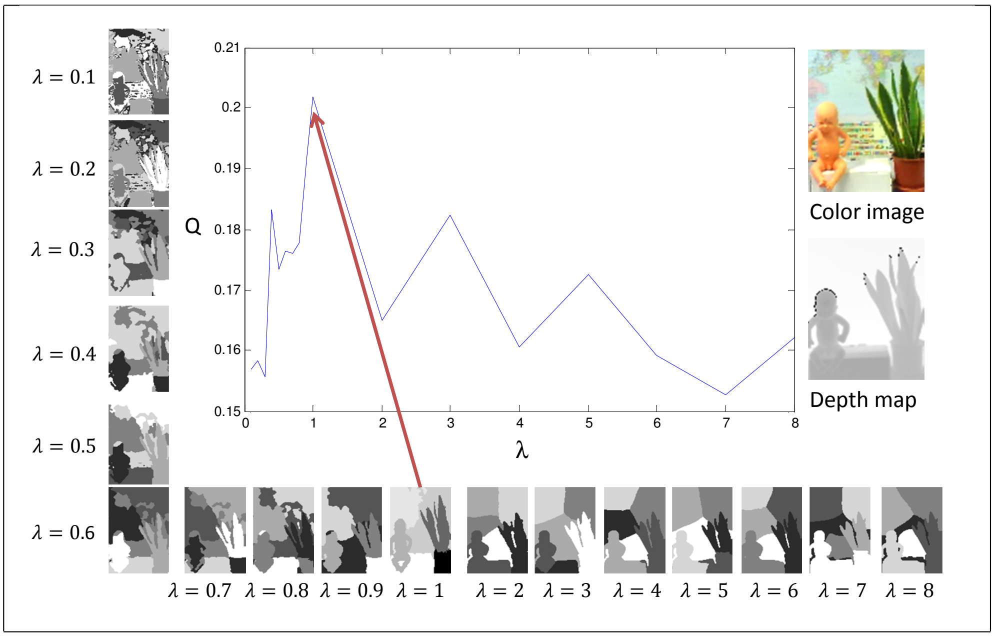


Figure 6.5: Values of Q versus λ and segmentation of the *baby and plant* scene for different values of λ

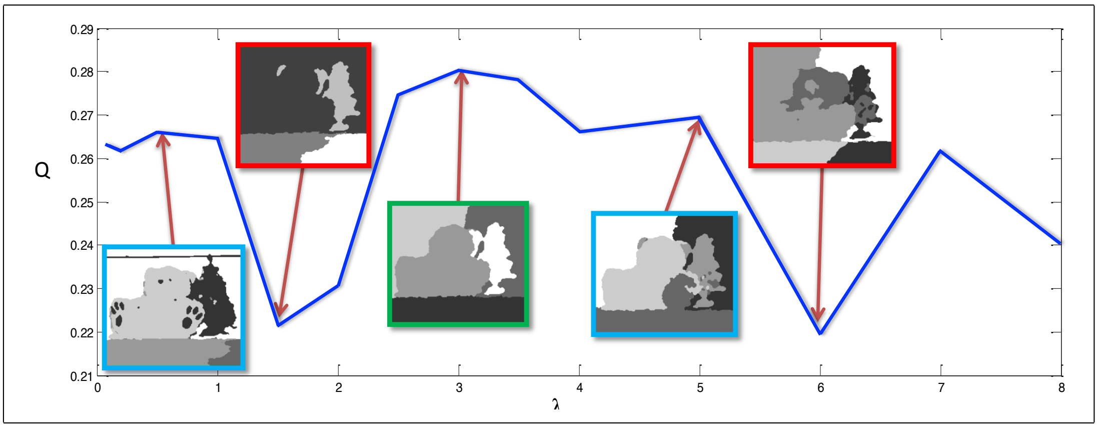


Figure 6.6: Values of Q versus λ and segmentation of the point cloud of the third row of Figure 6.18 for different values of λ. Note how the best segmentation (shown in green) is correctly recognized, good segmentations (in blue) correspond to high Q values and the bad segmentations (in red) to low Q values.

though the plots are quite different, in all the 3 cases the maximum of Q corresponds to the value of λ delivering the best segmentation result. It is finally worth noting that in spite this method requires to compute several segmentations, it can be easily managed within reasonable computation times by coarse to fine approaches. For instance a set of segmentations can be firstly performed on a subsampled dataset and then, once the optimal λ value is selected, the full resolution segmentation can be computed only for that value of λ. Furthermore in the case of video segmentation, since the optimal λ depends on the general scene properties, it could be computed on the first frame and then propagated to a set of subsequent frames.

6.5 Segmentation of stereo image pairs

Stereo vision algorithms are rather attractive for various reasons: there is a copious literature about them [86], they require an inexpensive setup and they use only a pair of images as input data, hence representing the next step in terms of acquisition complexity with respect to segmentation based on single images. Stereo vision data therefore represent a situation of special interest for the proposed segmentation approach. In this section an ad-hoc extension of the proposed method for this kind of data is proposed. It is worth noting though that the segmentation scheme introduced so far can already provide very good performance without the further extension of this section. This optional refinement allows to improve performance in the cases where two images and a depth-map are available.

As it is well known, a stereo vision system is constituted by two standard cameras that acquire two slightly different views of the same scene. If the stereo vision is calibrated, depth information can be estimated from the two views by one of the many stereo vision algorithms (see [12] for a comparison of state-of-the-art algorithms in this field). The segmentation method introduced so far can already be applied to the depth-map obtained from stereo vision and to one of the two images. However since in this case a second image of the same scene is also available, this section introduces a way to exploit it in order to further improve the segmentation results.

Lets denote with $\mathcal{L}(p_i)$ and $\mathcal{R}(p_i)$ the pair of rectified images and with $\mathcal{D}(p_i)$ the disparity map estimated from them (relative to the left view). Without loss of generality assume that the target is the segmentation of the scene as seen from the left image

$\mathcal{L}(p_i)$. The disparity map can be used to locate for each pixel of the left image the corresponding one in the right image, except for the pixels that are visible only in the left view (because of occlusions or because they are out of the right frame) or the pixels without a disparity value because of the limitations of the adopted stereo vision algorithm. Hence it is worth defining image $\mathcal{R}_w$ as follows:

$$\mathcal{R}_w(p_i) = \begin{cases} \mathcal{R}(p_i - \mathcal{D}(p_i)) & \text{if } \mathcal{D}(p_i) \text{ exists} \\ \mathcal{L}(p_i) & \text{if } \mathcal{D}(p_i) \text{ does not exist} \end{cases} \tag{6.19}$$

Image $\mathcal{R}_w(p_i)$ represents the right image warped to the viewpoint of the left one except for the points of the left image not visible in the right one. For these points the corresponding value in the left image is simply copied onto $\mathcal{R}_w(p_i)$. Figure 6.7d shows an example of such an image. The disparity map is related to the depth-map $\mathcal{Z}(p_i)$ through the well-known Equation (3.1) which in this notation can be written as $\mathcal{Z}(p_i) = (bf)/\mathcal{D}(p_i)$ where b is the baseline of the stereo vision setup and f focal length of the two cameras. The depth-map can then be used together with calibration information in order to compute the positions of the scene points in the 3D space. Therefore in the stereo case for each scene point p there is available:

- its color value in the left view $\mathcal{L}(p_i) = [L_l(p_i), a_l(p_i), b_l(p_i)]$
- its color value in the right view $\mathcal{R}_w(p_i) = [L_r(p_i), a_r(p_i), b_r(p_i)]$ (as previously said replaced by a copy of $\mathcal{L}(p_i)$ for the points not visible in the right view)
- its position in the 3D space $(x(p_i), y(p_i), z(p_i))$

As in Section 6.2 all the various components can be normalized by the corresponding standard deviations obtaining the three normalized vectors:

$$\begin{bmatrix} \bar{L}_l(p_i) \\ \bar{a}_l(p_i) \\ \bar{b}_l(p_i) \end{bmatrix} = \frac{1}{\sigma_{cl}} \begin{bmatrix} L_l(p_i) \\ a_l(p_i) \\ b_l(p_i) \end{bmatrix}$$

$$\begin{bmatrix} \bar{L}_r(p_i) \\ \bar{a}_r(p_i) \\ \bar{b}_r(p_i) \end{bmatrix} = \frac{1}{\sigma_{cr}} \begin{bmatrix} L_r(p_i) \\ a_r(p_i) \\ b_r(p_i) \end{bmatrix}$$

$$\begin{bmatrix} \bar{x}(p_i) \\ \bar{y}(p_i) \\ \bar{z}(p_i) \end{bmatrix} = \frac{1}{\sigma_g} \begin{bmatrix} x(p_i) \\ y(p_i) \\ z(p_i) \end{bmatrix}$$

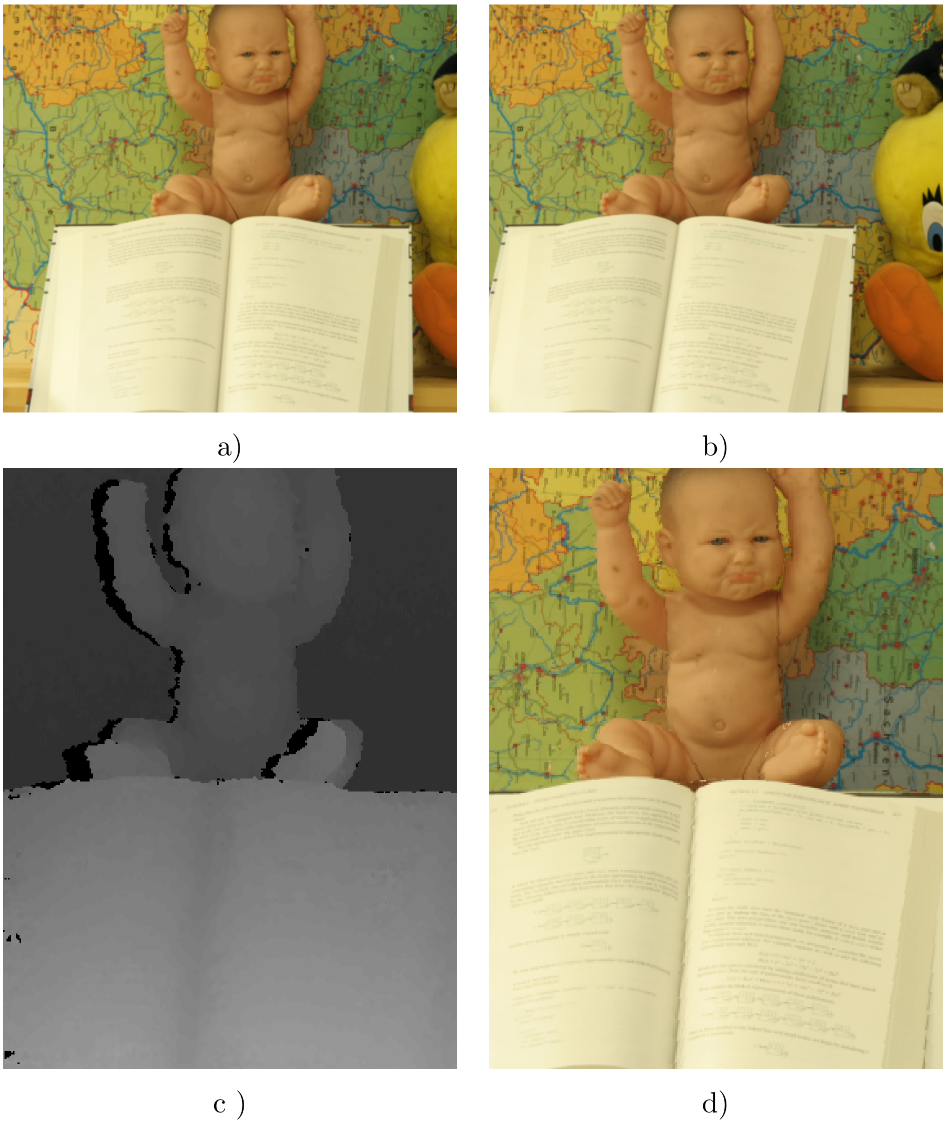
where the standard deviations σ_{Ll}, σ_{al} and σ_{bl} refer to the left view and σ_{Lr}, σ_{ar} and σ_{br} to the right one. Let $\sigma_{cl} = (\sigma_{Ll} + \sigma_{al} + \sigma_{bl})/3$ and $\sigma_{cr} = (\sigma_{Lr} + \sigma_{ar} + \sigma_{br})/3$ be the average standard deviations of color data for the left and right image respectively. The standard deviation of the geometry data is defined as in Section 6.2. From the above normalized geometry and color information vectors each scene point $p_i, i = 1, ..., N$ can be represented by a 9-dimensional vector representing its 3D position and its color in the two views naturally extending the representation of Section 6.2:

$$\mathbf{p}_i^f = \begin{bmatrix} \bar{L}_l(p_i) \\ \bar{a}_l(p_i) \\ \bar{b}_l(p_i) \\ \bar{L}_r(p_i) \\ \bar{a}_r(p_i) \\ \bar{b}_r(p_i) \\ \lambda\bar{x}(p_i) \\ \lambda\bar{y}(p_i) \\ \lambda\bar{z}(p_i) \end{bmatrix}, i = 1, ..., N \qquad (6.20)$$

This 9-dimensional vector can be used as input to the spectral clustering algorithm of Section 6.3 and used to segment the scene seen from the left image. In case the segmentation of both views was needed the same approach can be clearly adopted with the disparity map relative to the right view and by swapping the left and right images in the previous discussion. The advantage of the 9-dimensional representation is clearly motivated by the experimental results presented in Section 6.6.2.

6.6 Experimental Results

The performances of the proposed scene segmentation algorithm is verified on datasets representing different scenes, acquired with different technologies. This is purposely done in order to assess the effectiveness of the joint usage of color and geometry for scene segmentation, independently of the specific 3D data types and of the used acquisition tools. In particular the considered scenes are acquired by: a trinocular system made by a ToF camera and two standard cameras; a standard 2-views stereo vision system; a Microsoft Kinect sensor [9] and by Microsoft Phothosynth [10], *i.e.*, an unstructured scene reconstruction system.



a)

b)

c)

d)

Figure 6.7: Input data for the segmentation of stereoscopic pairs: a) left view; b) right view; c) disparity relative to the left view (disparity values have been stretched in order to improve the readability); d) detail of the right view warped to the left viewpoint. Note how occlusions in the warped view were filled by copying data from the left view. Some small artifacts noticeable in the figure are due to the errors in the disparity estimation (in this case estimated by the method of [63]).

6.6.1 Results on the trinocular system data

The setup presented in Chapter 5 can be used as a single system for acquiring both geometry and color information. The input data is obtained by taking all the 3D points acquired by the ToF camera and by appending to them the color information of the corresponding pixels obtained from the images of the two cameras. It is preferable to deploy two RGB cameras rather than only one in order to alleviate the occlusion problems. The proposed segmentation algorithm is tested on several scenes and compared with scene segmentation based on geometry or color information only obtained both by using our method and two state-of-the-art segmentation algorithms (*i.e.*, the graph-based method of Felzenszwalb et al. [52] and the mean-shift algorithm of [3]). The results of Figure 6.8 clearly show the effectiveness of the proposed method. The scenes shown in the figure contain good examples of common issues making non-trivial scene segmentation, namely issues due to the background color articulation and to the complexity of the scene geometry (as in the case of the plant of the second and third rows of the figure). The first two columns of Figure 6.8 show the color and geometry information relative to three different scenes (one for each row). These data have been used as input for three different segmentation methods (namely NNCSC, [52] and [3]) using either color information only or geometry information only and the corresponding results are shown in rows from 3 to 8. Finally the rightmost column shows the results of the proposed segmentation technique based on the fusion of color and geometry information. Color based segmentation exhibits various problems, *e.g.*, the space between the arms is not so clearly recognizable in the color segmentation results of the first row of Figure 6.8. In the scene of the first row of Figure 6.8 segmentation based on geometry information only gives better results, although not completely satisfactory (*e.g.* [3] provides the best results, indeed it is the only method that recognizes the two regions but the separation is not as accurate as for the proposed method). The proposed technique fusing color and geometry clearly performs better than the compared state-of-the-art algorithms. For instance in the case of the scene of the first row of Figure 6.8 it is the only method that accurately separates the baby from the white box behind it. The second and third rows of Figure 6.8 confirm that the proposed scene segmentation method allows for a very good segmentation of both the plant and the vase which are very difficult subjects to segment on the basis of either color or geometry only (*e.g.*, the proposed method is the only one capable to correctly extract the complete baby shape in the third row experiment).

It is fair to recall that the proposed technique incorporates NNCSC as clustering method. The usage of either k-means or of mean-shift as clustering method would give poorer results as shows the comparison of the results of the first row of Figure 6.8 with the ones of Figure 6.3. Figure 6.9 refers to the *baby and plant* scene (the one in the last row of Figure 6.8) and offers an extensive comparison between the results of different clustering techniques, namely it compares the proposed method based either on NNCSC, k-means or mean-shift and the techniques of [52] and [3]. Each row corresponds to a different method, while the different columns show the results on color only, on geometry only, and on the fusion of color and geometry. The results of row 4 and 5, obtained by the state-of-the-art image segmentation methods of [52] and [3] on either color only or geometry only information, demonstrate the effectiveness of the fusion of color and geometry by the proposed method. It is also worth noting how the proposed approach implemented with simpler clustering schemes would have a performance inferior to the one obtained by using NNCSC even if applied to color and geometry together.

Figure 6.10 refers instead to the segmentation of a person. It can be shown that a human shape is perfectly identified by the proposed method (Figure 6.10e), in contrast to the very bad result obtained by color information only, and to the one obtained by geometry only, that presents artefacts in the lower part of the body (*e.g.*, feet). This is a good example of a typical issue of segmentation based on geometry only. Geometry information turns out well suited to separate objects and people from the background, but not to separate different objects in touch with each other. At the same time color segmentation is prone to be mislead by complex texture patterns, such as the texture on the person's shirt. By suitably fusing the two clues it is possible to solve both issues at the same time.

The execution time of the current MATLAB implementation of the proposed segmentation algorithm was less than 0.5 seconds on all the analysed scenes.

6.6.2 Results on stereo vision data

The proposed scene segmentation method was also tested on data obtained from a stereo vision system (for these results geometry was recovered using the method of

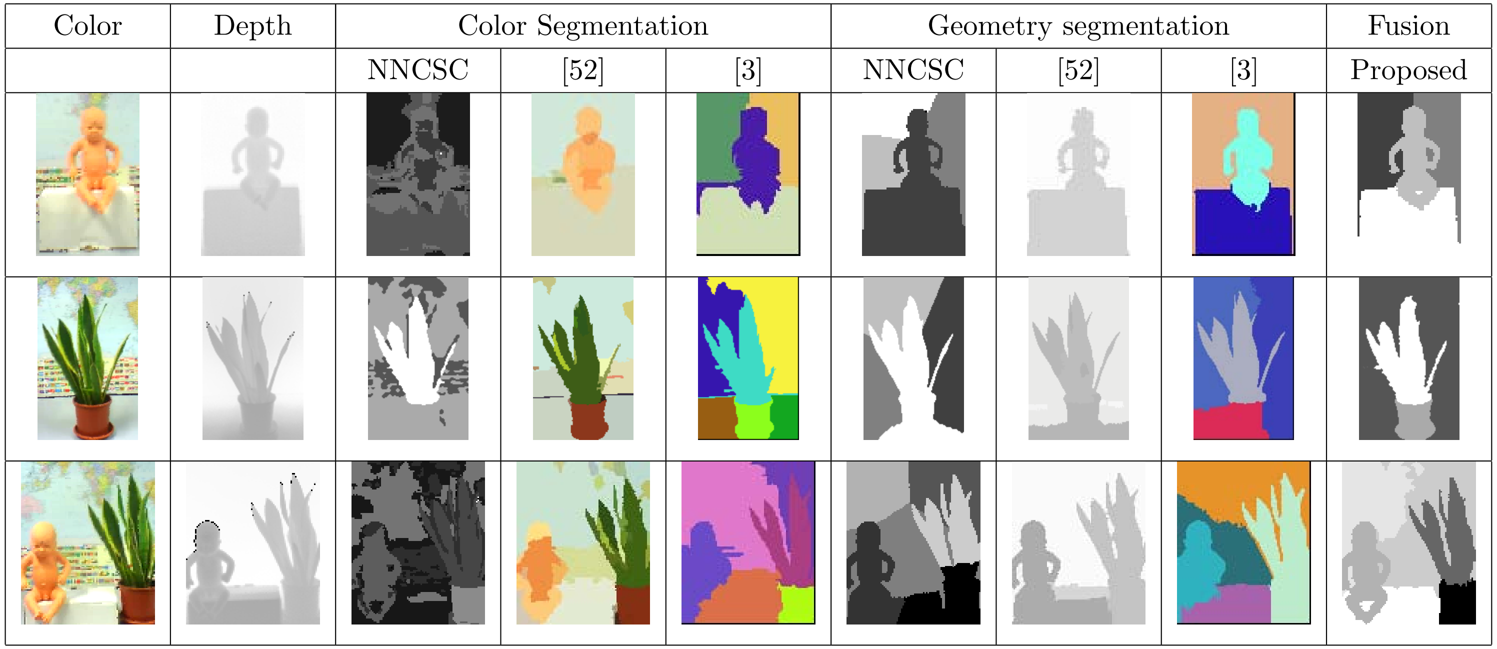


Figure 6.8: Segmentation of datasets acquired by the trinocular setup: (first row) *baby* scene, (second row) *plant* scene, (third row) *baby and plant* scene. The figure shows the results of the proposed method (rightmost column) using only color, only depth and by fusing the two clues. The figure also reports the results of two state-of-the-art methods (*i.e.* [52] and [3]) applied to color or geometry only.

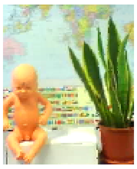
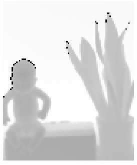
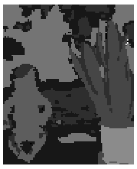
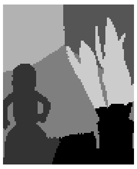
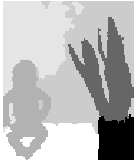
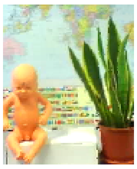
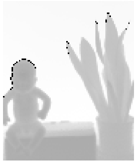
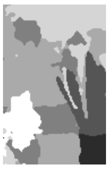
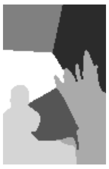
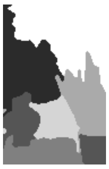
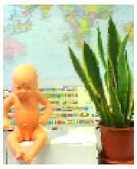
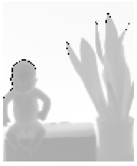
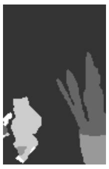
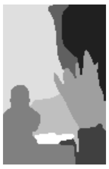
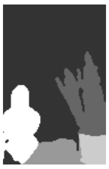
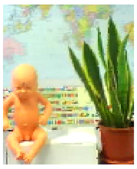
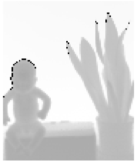
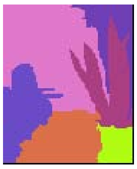
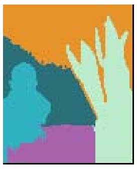
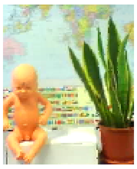
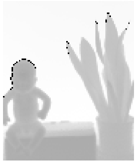
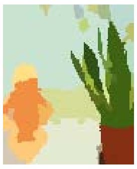
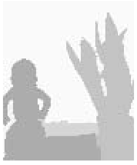
Method	Color image	Depth map	Color segm.	Geom. segm.	Fusion
Proposed method (Spectral Clust.)					
K-means clustering					
Mean-shift clustering					
Felzen. et Al.[52]					✖
Edison [3]					✖

Figure 6.9: Segmentation of the *baby and plant* scene using different segmentation algorithms on color, geometry and the fusion of color and geometry by the proposed approach.

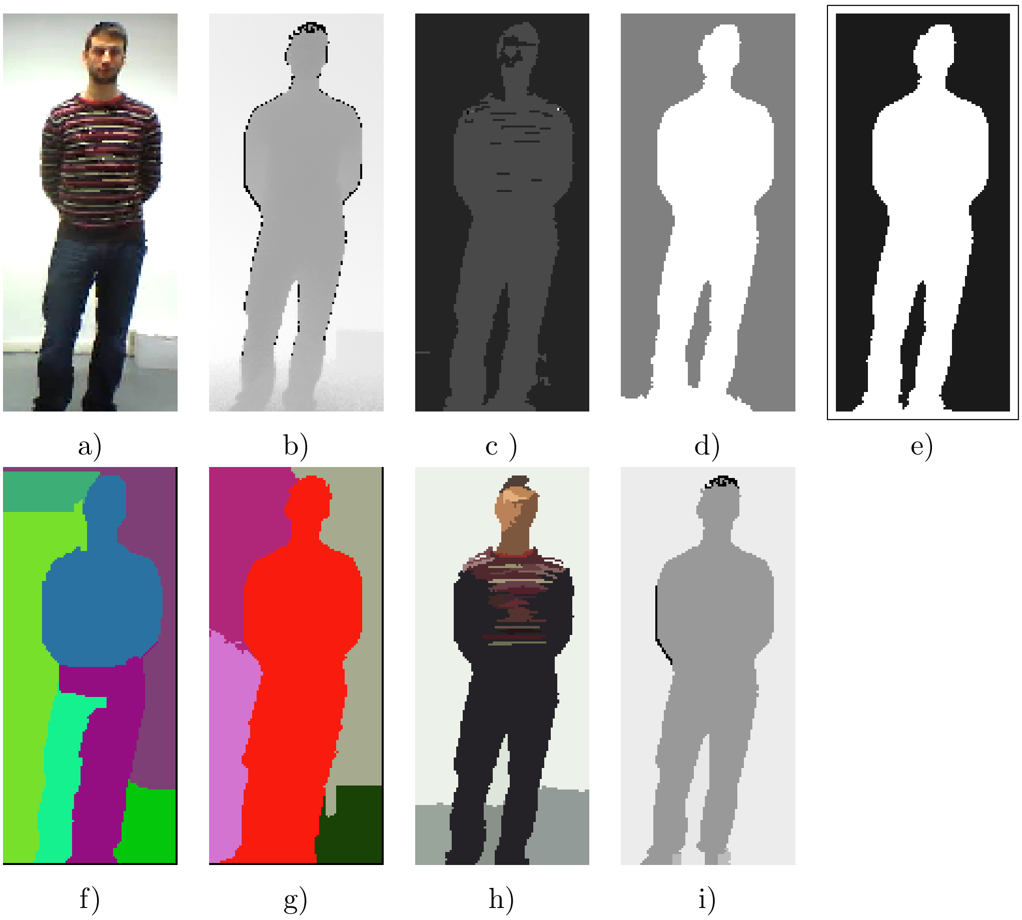


Figure 6.10: Segmentation of the datasets acquired by the trinocular setup on a person scene. The figure shows: a) color image; b) corresponding depth-map; c) segmentation on the basis of color information only; d) segmentation on the basis of geometry only; e) segmentation based on the proposed method, fusing geometry and color; f) segmentation obtained by applying [52] to color information; g) segmentation obtained by applying [52] to geometry information; h) segmentation obtained by applying [3] to color information; i) segmentation obtained by applying [3] to geometry information.

[63]). Our segmentation algorithm was tested on data from the Middlebury [12] stereo vision repository which is a very commonly used benchmark for stereo vision. Figure 6.11a and Figure 6.11b show the input data of the *aloe* scene of [12]. This is a quite challenging scene due to the heavily texturized background and to the complex shape of the plant. Figure 6.11c shows the result of the segmentation by the proposed method applied to one of the two views together with depth data. The results are already quite good: most of the leaves are recognized and the vase is correctly separated from the plant. However some artifacts are still present, *e.g.*, the artifacts on the right side of the vase due to the dark background or the ones on the upper right leaf. Figure 6.11d shows the benefits of the approach described in Section 6.5 that exploits also the second color view. Segmentation accuracy is improved (*e.g.*, the upper right leaf is correctly detected and the artefact on the right of the vase disappears). However some artifacts due to missing values in the depth data computed by [63] are still visible (*e.g.* on the side of some leafs). Figure 6.11e shows the results obtained by also applying an occlusion handling scheme [44], note how the artifacts due to missing depth data disappear. Figure 6.11f shows the results of [26], that also jointly exploits depth and color, while the Figures from 6.11g to 6.11l show the results of state-of-the-art segmentation algorithms working on either color only or geometry only. The proposed method (the results of the complete scheme are the ones of Figure 6.11e) clearly outperforms the other approaches.

Figure 6.12 refers instead to the *baby2* scene of the Middlebury repository. Again the proposed approach (Figure 6.12e) outperforms the other approaches shown in the Figures from 6.12f to 6.12l. In this case the results of the proposed approach are already very good with a single color view, however the exploitation of the second color view allows to get rid of a couple of minor remaining artefacts.

The performances of the proposed approach are also compared with other recent segmentation schemes jointly exploiting color and depth information. Figure 6.13 shows a comparison[1] between the proposed scheme and the methods of [34] and [26] on two scenes from the Middlebury dataset. The proposed method is the only one that in both situations correctly recognizes all the three main regions of the scene (*i.e.* vase, plant and background in the first and baby, box and background in the second). The

[1] The figures with the results of [34] have been taken from their paper while the method of [26] has been implemented following the description on the paper.

method of [34] can correctly recognize the foreground region shape but it cannot divide the objects on the basis of color information (it appears a bit biased towards depth data), while the method of [26] produces some artifacts (*e.g.* on the left side of the baby or close to the plant leaves), even if it is able to distinguish the baby from the box. Furthermore note how the proposed method allows to automatically balance the two clues, while the method of [26] requires a manual parameter tuning in order to obtain a good segmentation.

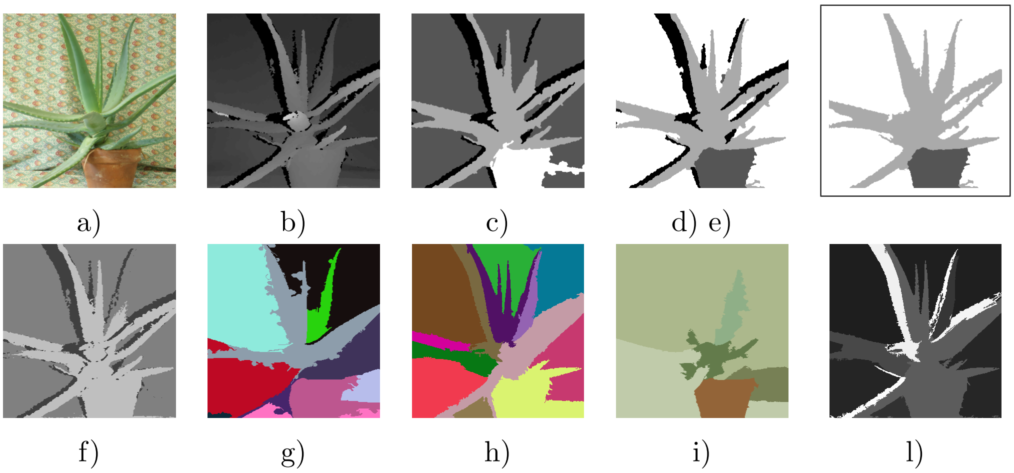


Figure 6.11: Segmentation of the *aloe* scene from the Middlebury dataset: a) color image; b) corresponding disparity map (disparity values have been stretched in order to improve the readability of the printed picture); c) segmentation based on the proposed method exploiting geometry and one of the color views; d) segmentation based on the proposed method exploiting both color views and geometry as described in Section 6.5; e) segmentation based on the proposed method exploiting both color views and geometry and also the occlusion handling scheme of [44]; f) segmentation performed by [26] that jointly exploits color and depth data; g) segmentation performed by [52] on the basis of color information only; h) segmentation performed by [52] on the basis of depth information only; i) segmentation performed by [3] on the basis of color information only; l) segmentation performed by [3] on the basis of depth information only.

6.6.3 Results on Kinect data

Nowadays, scene descriptions accounting for both geometry and color can be readily and inexpensively obtained also by cheap mass market devices such as the Microsoft Kinect [9]. In fact, the Kinect sensor includes both an active system that captures a

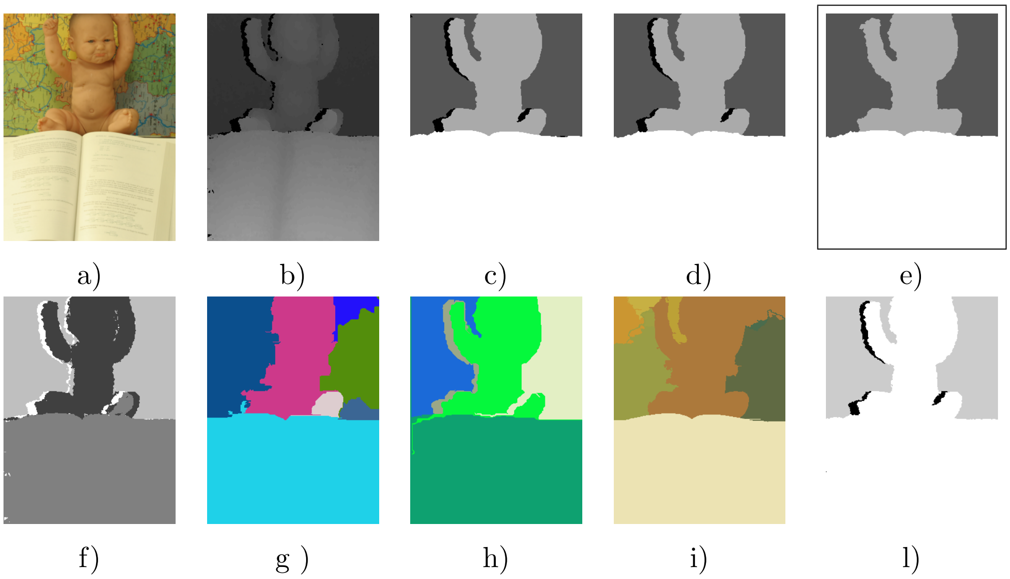


Figure 6.12: Segmentation of the *baby2* scene from the Middlebury dataset: a) color image; b) corresponding disparity map (disparity values have been stretched in order to improve the readability of the printed picture); c) segmentation based on the proposed method exploiting geometry and only one of the color views; d) segmentation based on the proposed method exploiting both color views and geometry as described in Section 6.5; e) segmentation based on the proposed method exploiting both color views and geometry and also the occlusion handling scheme of [44]; f) segmentation performed by [26] that jointly exploits color and depth data; g) segmentation performed by [52] on the basis of color information only; h) segmentation performed by [52] on the basis of depth information only; i) segmentation performed by [3] on the basis of color information only; l) segmentation performed by [3] on the basis of depth information only.

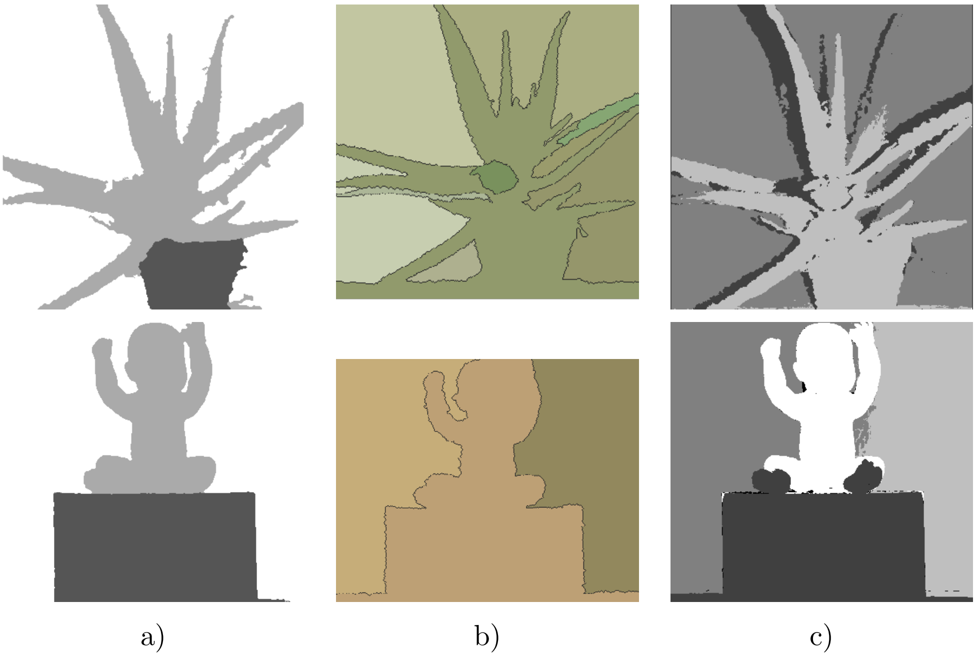


Figure 6.13: Comparison of different segmentation methods based on the joint use of depth and color information on the *aloe* scene (*first row*) and on the *baby1* scene (*second row*): a) Proposed method; b) Calderero and Marques [34]; c) Bleiweiss and Werman [26].

real time description of the scene geometry and a color camera. The wide availability and low cost of such sensors open a wide application scenario to the proposed segmentation framework since it eliminates the need of expensive 3D acquisition devices or of computationally complex state-of-the-art stereo algorithms.

In order to take advantage of both the geometry and the color acquired by the Kinect in a unique framework, it is firstly necessary to calibrate its depth sensor with the color camera. A first possibility is to perform a standard stereo camera calibration with OpenCV [13] on the color images acquired by the color camera and on the amplitude image acquired by the depth camera (with the IR projector obscured). The proposed segmentation algorithm can then be applied to the Kinect data as shown by the results of Figure 6.14. It is worth noting that the overall scene segmentation is correct, but there are some errors near depth discontinuities. Such errors are due to the artefacts present in the depth data acquired by the Kinect sensor (*i.e.*, the acquired depth and color edges are not precisely aligned as clearly visible in Figure 6.14a and Figure 6.14b).

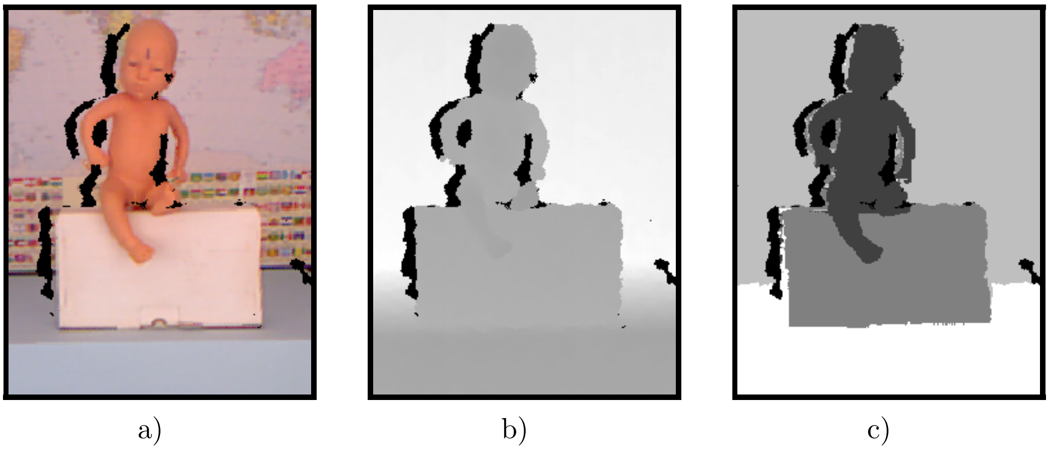
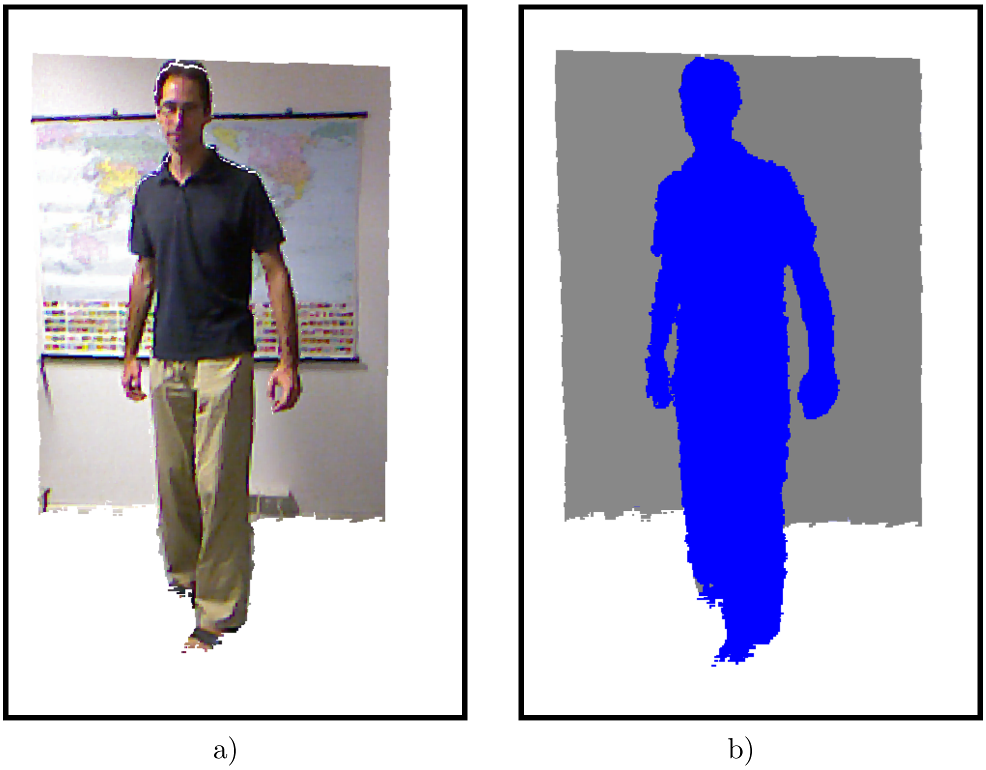


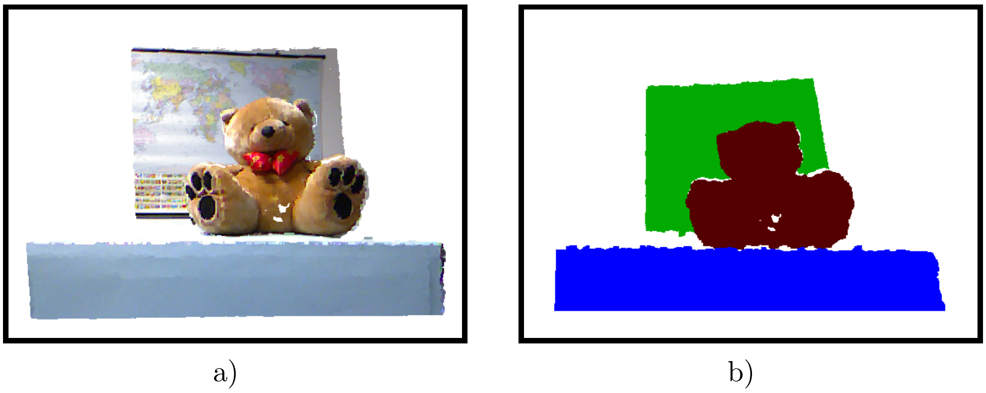
Figure 6.14: Segmentation of the *baby* scene acquired with a Kinect sensor: a) color image, b) depth image, c) segmented image.

A second possibility offered by the freely available OpenNI [14] framework is to directly acquire a colored point cloud. Figure 6.15 and Figure 6.16 show a couple of point clouds acquired in this way and the corresponding segmentations. Again the results are very good and the objects are correctly separated from the background (even the part of the teddy bear that touches the table is correctly separated from the table itself).



a) b)

Figure 6.15: Segmentation of a person scene acquired with a Kinect sensor: a) point cloud acquired by the Kinect sensor, b) segmentation of the point cloud.



a) b)

Figure 6.16: Segmentation of a teddy bear acquired with a Kinect sensor: a) point cloud acquired by the Kinect sensor, b) segmentation of the point cloud.

6.6.4 Results from Photosynth data

The acquisition systems of Sections 6.6.1 and 6.6.2 are classical tools capable to acquire dense representations of both geometry and color of a scene in terms of an image and the corresponding depth-map. An unstructured 3D scene reconstruction tool like Microsoft Photosynth [10] is rather attractive not only because it is a free tool but also because it just requires to shoot a number of uncalibrated standard pictures of the scene. Photosynth can now be used even on mobile phones and is probably the only way today available for obtaining 3D data by mobile phones. The major limitation of Photosynth is that it is only able to provide a sparse representation of the scene geometry and color since the geometry is estimated only for salient features-point. Color information can be associated to such salient points. The main characteristic of a salient region is that it is markedly different from the rest of the scene. Therefore, grouping a set of salient points means grouping points that by construction and assumption are significantly different from each other. This characteristic of the acquisition system is by itself rather problematic. Another challenge for the segmentation is given by the sparsity of the obtained point cloud. Another important characteristic of the data is that the estimated scene geometry is defined up to an arbitrary scale factor. We tested our algorithm on the scene of Figure 6.17, obtained by Photosynth. Figure 6.17b shows the resulting segmentation (each color in the image corresponds to a scene segment). In light of the complexity of the point-cloud, and of the difficulties inherent to this type of data as observed above the results can be considered remarkably good.

6.7 Comparison of the considered imaging systems for scene segmentation purposes

As shown in the experimental results the proposed segmentation scheme can be applied to the data coming from different 3D acquisition systems. Two interesting questions that may arise at this point concern how the segmentation accuracy depends on the employed acquisition system and which is the best imaging system for segmentation purposes. In order to give a first answer to these questions a set of different scenes is acquired with 3 different imaging systems, *i.e.* the trinocular system described in Section 6.6.1, a Kinect camera and a stereo vision system exploiting the algorithm of [63]. The acquired data are segmented exploiting the method proposed in this chapter

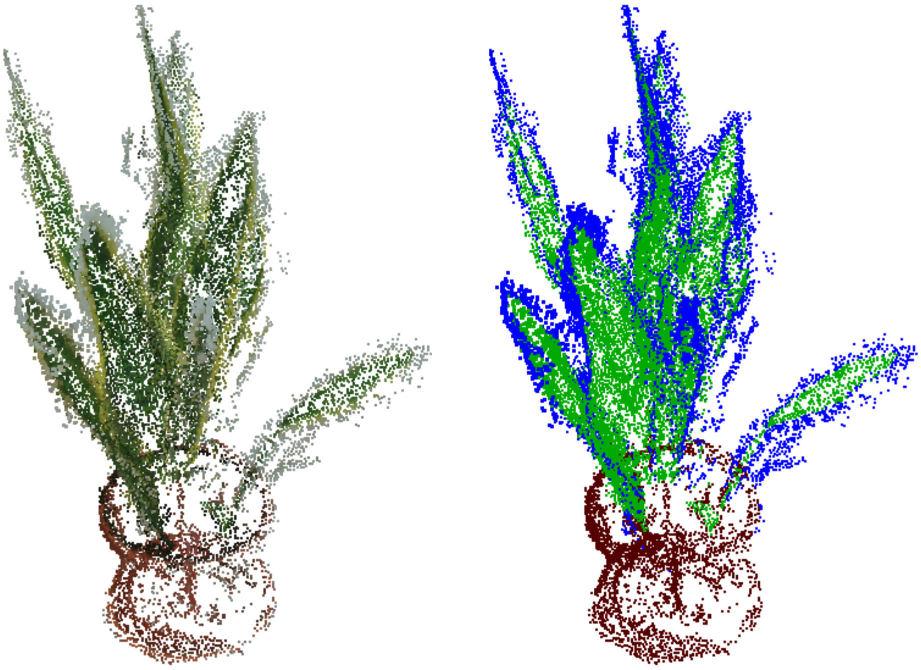


Figure 6.17: Segmentation of the scene *plant* acquired with Photosynth [10]: a) acquired scene, b) scene segmented by the proposed method jointly exploiting geometry and color. *(Best viewed in colors)*

and Figure 6.18 shows the obtained segmentations. Each of the five rows of the Figure 6.18 corresponds to a different scene (shown in the first column), while each of the last three columns corresponds to a different acquisition system. It is clear that the trinocular setup (column b) gives the best results. This is mostly due to two reasons: firstly there are not occluded areas due to the fact that the ToF camera does not suffer from this issue, secondly the depth data are more accurate than the data produced by the other acquisition devices. Note in particular that edge localization is more precise than the ones of the other devices. Unfortunately it is also the most expensive of the three systems. In spite the Kinect is a much cheaper solution, it can be effectively exploited for joint color and depth segmentation. Even from the data of this cheap device it is possible to recognize all the main objects in the framed scenes (as shown by the images in column c). Probably the biggest limit of the Kinect data is the edge localization. It suffers both from the edge artifacts typical of the depth data acquired by the Kinect and from the limited accuracy of the calibration between the color and the depth camera. Note how we used the internal calibration provided by the Kinect that is not as precise as the one we performed for the trinocular setup. The higher spatial resolution of the Kinect with respect to that of ToF cameras is of little use for segmentation purposes because of its poor edge localization. Stereo vision (column d) gives the worse results mostly because of the artifacts in the provided depth data and of the missing depth samples due to occlusions. This is an issue also in the case of Kinect, but the number of samples without a depth value is much smaller in this case than in the case of stereo vision systems. Artifacts in the computed depth-maps due to uniformly textured regions also affects the segmentation, in particular on the background of the considered scenes. The results of stereo systems shown in column d are also not so good as the ones of Section 6.6.2. This is due to the fact that the used stereo vision algorithm (but it is a problem common to many stereo techniques) performs very well on heavily textured scenes built *ad-hoc* for stereo vision testing, *e.g.* the ones of the Middlebury dataset, but not as well with real scenes. However stereo setups are also very inexpensive and do not require active lighting. They can also be used for the acquisition of large-scale and outdoor scenes while both the Kinect and the ToF camera can only measure distances up to a few meters and essentially cannot work outdoor since they are heavily affected by sunlight. As summarized by Table 6.1, each of the considered acquisition systems has its own advantages and disadvantages and the choice of the proper setup should be done on the basis of the target application.

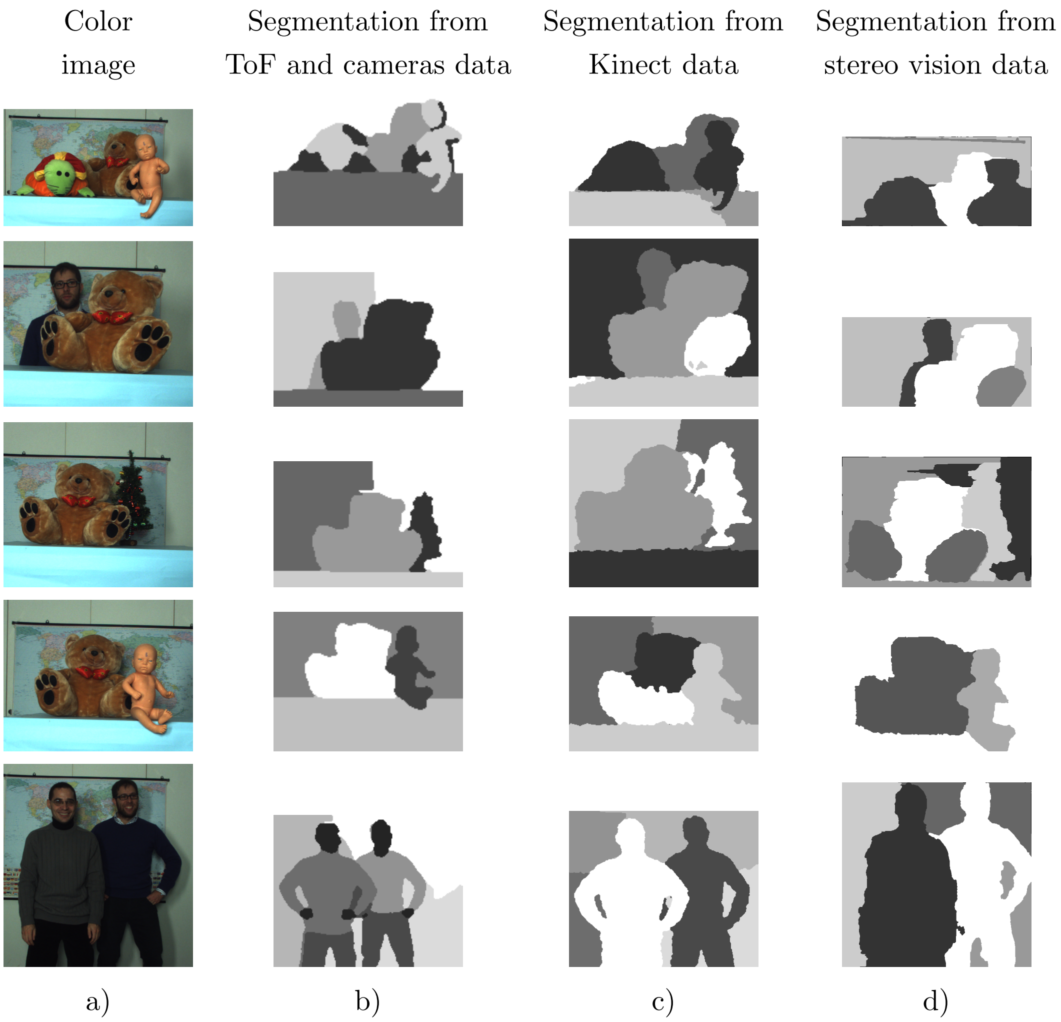

Color image
Segmentation from ToF and cameras data
Segmentation from Kinect data
Segmentation from stereo vision data
a)
b)
c)
d)

Figure 6.18: Segmentation of some samples scenes exploiting depth data coming from different acquisition systems: a) color image of the scene; b) segmentation from the ToF camera data and the color images provided by the trinocular setup; c) segmentation from the Kinect data; d) Segmentation from the stereo vision data.

	Trinocular Setup (ToF + cameras)	Microsoft Kinect	Stereo vision
Edge localization	Good	Poor	Poor
Resolution	Low	Medium	High
Missing depth values	Very few	A few	Yes
Outdoor scenes	No	No	Yes
Cost	High	Low	Low

Table 6.1: Comparison of the different acquisition setups.

7

Conclusion

This thesis presents in a unique body the mainstream of the research I carried out during my Ph.D. studies. The main focus on the thesis is the acquisition and the processing of data acquired by a ToF camera and by a stereo vision system. In particular such data carries information about the color and the three-dimensional geometry of the framed scene. Concerning the acquisition part, a description of how ToF cameras and stereo systems work is presented, as well as a characterization of the quality of the data acquired by the two systems. In particular, for ToF cameras a novel error model for the data acquisition process is proposed aiming at characterizing important effects, such as the finite dimension of the ToF camera pixels. Classical metrological quantities such as accuracy, precision and resolution are revisited for matricial depth information acquisition systems. Stereo systems and ToF cameras are characterized in terms of these metrics and their complementary in terms of them is experimentally supported. A probabilistic ToF and stereo data fusion have been proposed in order to obtain a system which improves the performance of the two subsystems. Such fusion method exploits the proposed data acquisition model for ToF cameras and advanced models for stereo vision systems in a MAP-MRF Bayesian framework. An application of three-dimensional geometry and color data acquired also by the proposed system in order to tackle the problem of scene segmentation is also presented. Several experimental results support all the proposed techniques. ■

Bibliography

[1] **CSEM**. http://www.csem.ch. 5

[2] **Delcom**. www.delcomproducts.com. 64

[3] **EDISON**. coewww.rutgers.edu/riul/research/code/EDISON. 57, 97, 98, 99, 100, 101, 103, 104

[4] **FBK**. http://www.fbk.eu. 5

[5] **Guide to the Expression of Uncertainty in Measurement**. 31

[6] **IEE**. http://www.iee.lu. 5

[7] **Intel Inc.** www.intel.com. 23

[8] **Mesa Imaging**. http://www.mesa-imaging.ch. 5, 33, 34, 59

[9] **Microsoft Kinect$^{\text{TM}}$**. http://www.xbox.com/en-US/kinect. 3, 79, 95, 103

[10] **Microsoft Photosynth$^{\text{TM}}$**. http://photosynth.net/. 79, 83, 95, 108, 109

[11] **Microsoft$^{®}$**. http://www.microsoft.com. 5

[12] **Middlebury Stereo Vision Dataset**. http://vision.middlebury.edu/stereo/. 93, 102

[13] **OpenCV**. http://opencv.willowgarage.com/wiki/. 38, 106

[14] **OpenNI**. http://www.openni.org/. 106

[15] **Panasonic D-Imager**. http://www.panasonic-electric-works.com. 5

[16] **PMD Technologies**. http://www.pmdtec.com/. 5

[17] **Point Grey Research, Inc., http://www.ptgrey.com/products/stereo.asp**. 23

[18] **Softkinetic**. http://www.softkinetic.com/. 5

[19] **STMicroelectronics**. http://www.cswww.st.com. 3

[20] **TYZX, Inc. , http://www.tyzx.com**. 23

[21] J. Besag. **On the statistical analysis of dirty pictures**. *Journal of the Royal Statistical Society*, **B-48**:259–302, 1986. 45, 63

[22] L. Bezze, C. Dal Mutto, P. Zanuttigh, F. Dominio, and Guido Maria Cortelazzo. **ToF Cameras and Microsoft Kinect Depth Sensor for Natural Gesture Interfaces**. In *ACM CHItaly*, 2011. 3

[23] S. Birchfield and C. Tomasi. **Depth Discontinuities by Pixel-to-Pixel Stereo**. *International Journal of Computer Vision*, **35**(3):269–293, December 1999. 57

[24] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007. 29, 64

[25] M. J. Black and A. Rangarajan. **On the unification of line processes, outlier rejection, and robust statistics with applications in early vision**. *International Journal of Computer Vision*, **19**:57–91, 1996. 61

[26] A. Bleiweiss and M. Werman. **Fusing time-of-flight depth and color for real-time segmentation and tracking**. In *Proceedings of DAGM 2009 Workshop on Dynamic 3D Imaging*, pages 58–69, 2009. 80, 83, 102, 103, 104, 105

[27] M. Bleyer, C. Rother, P. Kohli, D. Scharstein, and S. Sinha. **Object Stereo- Joint Stereo Matching and Object Segmentation**. In *Conference on Computer Vision and Pattern Recognition*, June 2011. 80

[28] Y. Boykov and V. Kolmogorov. **An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**:359–374, 2001. 56, 58, 59

[29] Y. Boykov, O. Veksler, and R. Zabih. **Fast Approximate Energy Minimization via Graph Cuts**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**:1222–1239, 2001. 45, 63

[30] G. Bradski and A. Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library.* O'Reilly, 2008. 24, 25

[31] M. Burns and G.W. Roberts. *Mixed-Signal IC Test and Measurement.* The Oxford Series in Electrical and Computer Engineering. Oxford University Press, 2001. 31

[32] B. Buttgen, T. Oggier, M. Lehmann, R. Kaufmann, and F. Lustenberger. **CCD/CMOS lock-in pixel for range imaging: Challenges, limitations and state-of-the-art**. In *1st range imaging research day*, 2005. 7, 9, 10, 11, 12, 13

[33] B. Buttgen and P. Seitz. **Robust Optical Time-of-Flight Range Imaging Based on Smart Pixel Structures**. *IEEE Transactions on Circuits and Systems I*, **55**(6):1512 –1525, july 2008. 8, 11, 12

[34] F. Calderero and F. Marques. **Hierarchical fusion of color and depth information at partition level by cooperative region merging**. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing ICASSP 2009*, pages 973–976, 2009. 80, 102, 103, 105

[35] D. Claus and A. W. Fitzgibbon. **A Rational Function Lens Distortion Model for General Cameras**. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005. 16

[36] D. Comaniciu and P. Meer. **Mean shift: a robust approach toward feature space analysis**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002. 79

[37] I. J. Cox, S. L. Hingorani, S. B. Rao, and B. M. Maggs. **A maximum likelihood stereo algorithm**. In *Computer Vision and Image Understanding*, **63**, pages 542–567, 1996. 30

[38] F.C. Crow. **Summed-area tables for texture mapping**. In *Proceedings of the 11th annual conference on Computer graphics and interactive techniques, SIGGRAPH*, 1984. 28

[39] C. Dal Mutto, F. Dominio, P. Zanuttigh, and Guido M. Cortelazzo. **Hand Gesture Recognition for 3D Interfaces**. In *STreaming Day*, 2011. 3

[40] C. Dal Mutto, F. Dominio, P. Zanuttigh, and S. Mattoccia. *Current Advancements in Stereo Vision*, chapter Stereo Vision and Scene Segmentation. Intech, 2012. 3

[41] C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo. **Accurate 3D Reconstruction by Stereo and ToF Data Fusion**. In *Proceedings of GTTI meeting*, Brescia, Italy, June 2010. 2, 3

[42] C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo. **A Probabilistic Approach to ToF and Stereo Data Fusion**. In *Proceedings of 3DPVT*, Paris, France, May 2010. 2, 3, 44, 45, 47, 51, 58, 75, 77

[43] C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo. **Scene Segmentation by Color and Depth Information and its Application**. In *STreaming Day*, Udine, Italy, September 2010. 3

[44] C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo. **Fusion of Geometry and Color Information for Scene Segmentation**. *IEEE Journal of Selected Topics in Signal Processing*, September 2012. 3, 4, 87, 102, 103, 104

[45] C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo. *Time-of-Flight Cameras and Microsoft Kinect*. Springer Briefs, 2012. 1, 2, 3, 4, 9, 44, 48

[46] C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo. *ToF Range-Imaging Cameras*, chapter ToF Cameras and Stereo Systems: Comparisons and Data Fusion. Springer, 2013. 3, 4

[47] C. Dal Mutto, P. Zanuttigh, and G.M. Cortelazzo. **Scene segmentation assisted by stereo vision**. In *Proceedings of 3DIMPVT 2011*, Hangzhou, China, May 2011. 3

[48] C. Dal Mutto, P. Zanuttigh, S. Mattoccia, and G. M. Cortelazzo. **Locally Consistent ToF and Stereo Data Fusion**. In Springer, editor, *Proceedings of 2nd Workshop on Consumer Depth Cameras for Computer Vision*, Florence, Italy, 2012. 3, 44, 45

[49] J. Davis, D. Nehab, R. Ramamoorthi, and S. Rusinkiewicz. **Spacetime stereo: A unifying framework for depth from triangulation**. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003. 38, 65

[50] J. Diebel and S. Thrun. **An Application of Markov Random Fields to Range Sensing**. In *Proceedings of Conference on Neural Information Processing Systems (NIPS)*, Cambridge, MA, 2005. MIT Press. 44

[51] J. Dolson, J. Baek, C. Plagemann, and S. Thrun. **Upsampling range data in dynamic environments**. In *IEEE Conference on Computer Vision and Pattern Recognition*, **0**, pages 1141–1148. IEEE Computer Society, 2010. 44

[52] P.F. Felzenszwalb and D.P. Huttenlocher. **Efficient Graph-Based Image Segmentation**. *International Journal of Computer Vision*, 2004. 79, 97, 98, 99, 100, 101, 103, 104

[53] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. **Spectral grouping using the Nyström method**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004. 86

[54] A. Frick, F. Kellner, B. Bartczak, and R. Koch. **Generation of 3D-TV LDV-content with Time-Of-Flight Camera**. In *Proc. of 3DTV Conf.*, 2009. 44

[55] A. Fusiello, E. Trucco, and A. Verri. **A compact algorithm for rectification of stereo pairs**. *Machine Vision and Applications*, **12**:16–22, 2000. 25

[56] V. Fusiello, A. Roberto and E. Trucco. **Symmetric Stereo with Multiple Windowing**. *International Journal of Pattern Recognition and Artificial Intelligence*, **14**:1053–1066, 2000. 28

[57] V. Garro, C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo. **A Novel Interpolation Scheme for Range Data with Side Information**. In *CVMP*, pages 52 –60, nov. 2009. 44

[58] S. A. Gudmundsson, H. Aanaes, and R. Larsen. **Fusion of stereo vision and Time of Flight imaging for improved 3D estimation**. *Int. J. Intell. Syst. Technol. Appl.*, **5**:425–433, 2008. 44, 58

[59] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[60] M. Harville, G. Gordon, and J. Woodfill. **Foreground segmentation using adaptive mixture models in color and depth**. In *IEEE Workshop on Detection and Recognition of Events in Video, 2001. Proceedings.*, 2001. 80

[61] J. Heikkila and O. Silven. **A Four-step Camera Calibration Procedure with Implicit Image Correction**. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1997. 16

[62] C. E. Hernandez, G. Vogiatzis, and R. Cipolla. **Probabilistic visibility for multi-view stereo**. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 47

[63] H. Hirschmuller. **Stereo Processing by Semi-Global Matching and Mutual Information**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**:328–341, February 2008. 96, 102, 108

[64] H. Hirschmuller. **Stereo Processing by Semiglobal Matching and Mutual Information**. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**, pages 328–341, February 2008. 30

[65] MESA Imaging. **SR4000 user manual**. http://www.mesa-imaging.ch. 21

[66] I. S. Kweon K. J. Yoon. **Adaptive support-weight approach for correspondence search**. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**, pages 650–656, April 2006. 28

[67] T. Kahlmann and H. Ingensand. **Calibration and development for increased accuracy of 3D range imaging cameras**. *Journal of Applied Geodesy*, **2**:1–11, 2008. 9, 10, 33, 34, 44

[68] T. Kanade and M. Okutomi. **A stereo matching algorithm with an adaptive window: theory and experiment**. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **16**, pages 920 – 932, September 1994. 28

[69] Y. M. Kim, C. Theobald, J. Diebel, J. Kosecka, B. Miscusik, and S. Thrun. **Multi-view Image and ToF Sensor Fusion for Dense 3D Reconstruction**. In *Proc. of 3DIM Conf.*, 2009. 44

[70] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother. **Bi-layer segmentation of binocular stereo video**. In *IEEE Conference on Computer Vision and Pattern Recognition*, **2**, page 1186 vol. 2, june 2005. 80

[71] K. D. Kuhnert and M. Stommel. **Fusion of Stereo-Camera and PMD-Camera Data for Real-Time Suited Precise**, 2006. 44, 58

[72] L. Ladicky, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, W. Clocksin, and P. Torr. **Joint Optimisation for Object Class Segmentation and Dense Stereo Reconstruction**. **BMVC special award issue**, pages 1–12, 2011. 80

[73] R. Lange. *3D Time-Of-Flight distance measurement with custom solid-state image sensors in CMOS/CCD-technology*. PhD thesis, University of Siegen, 2000. 7, 11, 13, 44, 48

[74] J. Leens, S. Piérard, O. Barnich, M. Van Droogenbroeck, and J.-M. Wagner. **Combining Color, Depth, and Motion for Video Segmentation.** In *ICVS'09*, pages 104–113, 2009. 80

[75] M. Lehmann, R. Kaufmann, F. Lustenberger, B. Bttgen, and T. Oggier. **CCD/CMOS lock-in pixel for range imaging: Challenges, limitations and state-of-the-art**. *In CSEM, Swiss Center for Electronics and Microtechnology*, 2004. 44

[76] S. Z. Li. *Markov Random Field Modeling in Image Analysis.* Springer, New York, 3rd edition edition, 2009. 29, 60

[77] S. Mattoccia, S. Giardino, and S. Gambini. **Accurate and Efficient Cost Aggregation Strategy for Stereo Correspondence Based on Approximated Joint Bilateral Filtering**. In *ACCV*, 2009. 57, 58

[78] M.J. McDonnell. **Box-filtering techniques**. In *Computer Graphics and Image Processing*, **17**, September 1981. 28

[79] V. Mezaris, I. Kompatsiaris, and M.G. Strintzis. **Still Image Segmentation Tools for Object-based Multimedia Applications**. *International Journal of Pattern Recognition and Artificial Intelligence*, **18**(4):701–725, June 2004. 80

[80] F. Mufti and R. Mahony. **Statistical analysis of measurement processes for time-of flight cameras**. *Proceedings of SPIE the International Society for Optical Engineering*, 2009. Available from: `http://cat.inist.fr/?aModele=afficheN&cpsidt=22392582`. 7, 9, 10, 11

[81] R. Nair, F. Lenzen, S. Meister, H. Schaefer, C. Garbe, and D. Kondermann. **High Accuracy ToF and Stereo Sensor Fusion At Interactive Rates**. In Springer, editor, *Proceedings of 2nd Workshop on Consumer Depth Cameras for Computer Vision*, Florence, Italy, 2012. 44, 45

[82] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference.* Morgan Kaufmann Publishers Inc., 1988. 45, 63

[83] D. Piatti and F. Rinaudo. **SR-4000 and CamCube3.0 Time of Flight (ToF) Cameras: Tests and Comparison**. *Remote Sensing*, **4**(4):1069–1089, 2012. 44

[84] C. Rosenberger and K. Chehdi. **Genetic fusion: application to multi-components image segmentation**. In *ICASSP*, 2000. 87, 88, 90

[85] G. Sanguinetti, J. Laidler, and N. Lawrence. **A probabilistic approach to spectral clustering: Using KL divergence to find good clusters**. In

Pascal Statistics and Optimization of Clustering Workshop, London, UK, Jul 2005. 86

[86] D. Scharstein and R. Szeliski. **A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms**. *International Journal of Computer Vision*, 2001. 1, 27, 57, 93

[87] J. Shi and J. Malik. **Normalized Cuts and Image Segmentation**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000. 79, 81, 85, 86

[88] D. Stoppa and F. Remondino, editors. *TOF Range-Imaging Cameras*. Springer, 2012. 1, 3

[89] J. Sun, N. Zheng, and H. Shum. **Stereo Matching Using Belief Propagation**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**:787–800, 2003. 29, 56, 57, 58, 59

[90] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer, New York, 2010. 16, 24, 25, 26, 79

[91] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. **A Comparative Study of Energy Minimization Methods for Markov Random Fields with Smoothness-Based Priors**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**:1068 –1080, 2008. 29, 45, 63

[92] F. Tombari, S. Mattoccia, and L. Di Stefano. **Segmentation-based adaptive support for accurate stereo correspondence**. In *Proceedings of IEEE Pacific-Rim Symposium on Image and Video Technology*, Santiago, Chile, December 2007. 28

[93] C. Uriarte, B. Scholz-Reiter, S. Ramanandan, and D. Kraus. **Modeling Distance Nonlinearity in ToF Cameras and Correction Based on Integration Time Offsets**. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer Berlin / Heidelberg, 2011. 9

[94] M. Wainwright, T. Jaakkola, and A. Willsky. **MAP estimation via agreement on (hyper)trees: Message-passing and linear programming approaches**. *IEEE Transactions on Information Theory*, **51**:3697–3717, 2002. 45, 63

[95] M. Wallenberg, M. Felsberg, P. Forssen, and B. Dellen. **Channel coding for joint colour and depth segmentation**. In *Lecture Notes in Computer Science (Proceedings of the 33rd Annual Symposium of the German Association for Pattern Recognition)*, **6835**, pages 306–315. Springer, 2011. 80

[96] L. Wang, C. Zhang, R. Yang, and C. Zhang. **TofCut: Towards Robust Real-time Foreground Extraction using Time-of-flight Camera**. In *Proceedings of 3DPVT 2010*, Paris, France, May 2010. 80

[97] John W. Woods. *Multidimensional Signal, Image, And Video Processing And Coding*. Elsevier Inc., 2006. 53

[98] Q. Yang, K.H. Tan, B. Culbertson, and J. Apostolopoulos. **Fusion of Active and Passive Sensors for Fast 3D Capture**. In *Multimedia Signal Processing (MMSP), IEEE International Workshop on*, 2010. 44, 45, 58, 77

[99] Q. Yang, R. Yang, J. Davis, and D. Nister. **Spatial-Depth Super Resolution for Range Images**. In *IEEE Conference on Computer Vision and Pattern Recognition*, **0**, pages 1–8. IEEE Computer Society, 2007. 44, 58, 77

[100] H. Zhang, J.E. Fritts, and S.A. Goldman. **Image segmentation evaluation: A survey of unsupervised methods**. *CVIU*, 2008. 87

[101] L. Zhang, B. Curless, and S.M. Seitz. **Spacetime stereo: shape recovery for dynamic scenes**. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003. 38, 65

[102] J. Zhu, L. Wang, J. Gao, and R. Yang. **Spatial-Temporal Fusion for High Accuracy Depth Maps Using Dynamic MRFs**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**:899–909, 2010. 44, 45, 47, 48, 58, 59

[103] J. Zhu, L. Wang, R. Yang, and J. Davis. **Fusion of time-of-flight depth and stereo for high accuracy depth maps**. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 44, 45, 47, 48, 58, 59

[104] J. Zhu, L. Wang, R. Yang, J. E. Davis, and Z. Pan. **Reliability Fusion of Time-of-Flight Depth and Stereo Geometry for High Quality Depth Maps**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**:1400–1414, 2011. 44, 45, 47, 48, 77