Giulio Marin

Confidence Estimation of ToF and Stereo Data for 3D Data Fusion

Stima della Confidenza delle Misure Ottenute da Sensori ToF e da Sistemi Stereo per la Fusione di Dati 3D

Tesi di laurea magistrale

Advisor: Prof. Pietro Zanuttigh
Co-Advisor: Ph.D. Carlo Dal Mutto

University of Padova

School of Engineering

Department of Information Engineering

October 8, 2013

Al reparto di Chirurgia della mano e Microchirurgia

Azienda Ospedaliera "Santa Maria degli Angeli", Pordenone.

Abstract

Accurate depth maps estimation from Time-of-Flight (ToF) range cameras and stereo vision systems is an interesting and relevant problem in computer vision. These two families of imaging systems, considered alone have complementary characteristics. ToF cameras work well in conditions where stereo has some problems, e.g. with uniform bright scenes, on the other hand these sensors perform worse on textured and less reflective scenes for which stereo shows higher performance. This suggests that a fusion of the data acquired by the two subsystems might improve the quality of the acquired three-dimensional geometry information. A measure on the correctness of the two depth estimations is required to weight the two hypotheses in the fusion process.

This thesis focuses on the analysis of Time-of-Flight and stereo vision systems, with the goal of extracting reliable confidence measures associated to the computed depth maps. In the first part of this work, the two families of sensors are described and after an analysis on practical issues some confidence measures are provided. Then, a framework for 3D data fusion with confidence information is presented and evaluated over a dataset of real world data. Experimental results show that the proposed fusion approach outperforms the performance of the two systems alone.

Sommario

La stima accurata di mappe di profondità utilizzando sensori Time-of-Flight (ToF) e sistemi di visione stereo è un problema rilevante nella visione computazionale. Questi due sistemi di acquisizione se considerati da soli hanno caratteristiche complementari. Le camere ToF funzionano bene in condizioni dove i sistemi stereo hanno problemi, per esempio in scene con oggetti chiari e uniformi, mentre funzionano peggio in scene con materiali poco riflettenti e molta texture dove lo stereo presenta risultati migliori. Questo suggerisce che la fusione dei dati acquisiti dai due sistemi potrebbe migliorare la qualità delle informazioni tridimensionali della scena acquisita. Una misura della correttezza delle due mappe di profondità è richiesta, al fine di pesare le due ipotesi nel processo di fusione.

Questa tesi si focalizza sull'analisi dei sensori Time-of-Flight e dei sistemi stereo, con lo scopo di estrarre delle mappe di confidenza affidabili relative alle mappe di profondità calcolate. Nella prima parte di questo lavoro, dopo una descrizione delle due famiglie di sensori, vengono analizzati i principali problemi di questi sistemi e alcune mappe di confidenza. Successivamente viene presentato il framework sviluppato per la fusione di dati 3D con relative informazioni di confidenza e i risultati vengono valutati in scene reali. Gli esperimenti mostrano che la fusione delle mappe di profondità permette di ottenere risultati migliori rispetto ai due sistemi considerati separatamente.

Acknowledgements

I would like to gratefully acknowledge all of the people who somehow have helped me during my two last years.

Carlo Dal Mutto, who fired up my interest in the 3D world and who I admire to always have constructive suggestions. Imimtek guys to have believed in me. Prof. Guido Maria Cortelazzo who taught me the fundamentals of Computer Vision.

Annaclaudia Montanino that since we have met has always believed in me, for the runs together, and especially to always be present to cheer me up. Davide Del Testa and Marco Fraccaro, invaluable fellows which for two years have been committed to send me their notes and without whom it would not have been possible to do the exams.

I'm indebted to Pietro Zanuttigh who has let me spend three months at Imimtek while working on this thesis.

I'm also grateful to the many other friends, not mentioned before, who have always had a comforting word for me during my many misadventures.

A special thanks goes to my parents, Mauro and Rosanna, to my sister Sara and to Terenzio, who have always helped me. Finally I want to thanks Maria, who has always been source of inspiration to me and from which I learned a lot.

Padova, October 8, 2013 Giulio

Contents

List of Figures

Chapter 1

Introduction

Perception of the three-dimensional structure of the world around us is an apparently easy task for humans. Think of how accurately you can perceive the depth information by just looking around, a subtle combination of light, shadow and color that we perceive, allows our brain to create a depth map of the scene and to infer the 3D geometry with extreme accuracy.

Psychologists for many years have tried to understand how the visual system works, but a complete solution to this problem still remains elusive. At the same time, computer vision scientists have developed mathematical techniques to recover the 3D geometry of objects in imagery. They first mimic our vision system by combining pictures recorded by two adjacent cameras (also called stereo vision): exploiting the difference between them it is possible to gain a strong sense of depth. Then they introduced the usage of Time-of-Flight sensors to directly measure the distance of each portion of a scene. Recently, devices like Microsoft Kinect, claimed of being the "fastest selling consumer electronics device", have increased the usage of 3D data in different fields that go beyond the simple gaming and Natural user Interface. Applications are manifold and go from the most common machine vision and robotics to physical recovery and rehabilitation.

Stereo vision is a classical approach to acquire three-dimensional information of a scene and significant progress has been done during the last few decades, but results for real-time scenes acquisition are still inaccurate, occlusions and textureless regions being the fundamental problems. On the other hand, stereo systems are constituted by two standard cameras, therefore they are capable to deliver high resolution color images in different illumination scenarios, and potentially precise three-dimensional information of the scene.

Time-of-Flight cameras were introduced to solve the problems of stereo systems and indeed they are able to acquire three-dimensional scenes more robustly at the

cost of a higher price and a greater power consumption. Time-of-Flight cameras solve the occlusion problem of stereo systems, having a unique viewpoint, and the geometry of textureless object can be inferred by these sensors thanks to an InfraRed signal that measures the distance. Among the problems of these systems, the most crucial are the low resolution (almost 10 times lower compared to regular camera) and the strong influence of objects reflectivity and background illumination on the received signal.

Interestingly, these two families of acquisition systems have complementary limitations, therefore it is likely to believe that a combination of depth information from the two sensors could improve the overall depth quality.

In this thesis the working principles and practical issues of Time-of-Flight and stereo cameras are investigated, in order to derive some confidence maps associated to the estimated depth maps. With this information, the depth maps of the two sensors are synergically fused by means of an extended version of the Locally consistent techniques [19]. In particular, Chapter 2 provides a general description of the two families of sensors previously mentioned and of the overall imaging system. Chapter 3 and 4 are relative to Time-of-Flight camera and stereo vision system respectively, and are organized with the same structure: first the working principles and the relative practical issues are presented, then the algorithms used to estimate the depth maps are described, finally some confidence measures associated to the depth maps are derived. Chapter 5 describes how to combine the two computed depth maps and relative confidences to obtain a better depth estimation. Results of the fusion algorithm are discussed in Chapter 6. Finally, conclusions and possible future works are presented in Chapter 7.

1.1 Related works

The idea of combining ToF sensors with standard cameras has recently attracted many researchers in this field. In [43], the two sensors are combined in a Maximum A Posteriori fashion, and the prior probability comes from a Markov Random Field model. With the aid of stereo vision they improve the performance of Time-of-Flight sensors providing a combined geometric calibration. In [42] authors extend this approach by efficiently finding a maximum of the posterior probability using Loopy Belief Propagation to optimize a global energy function. The traditional spatial MRF has been extended to a dynamic scenario with temporal coherence.

In [39] the ToF depth is up-sampled using joint bilateral filtering, then an amplitude based confidence map is used together with a stereo confidence map

based on local features, to combine the respective cost volumes. The final depth map is obtained with a greedy algorithm.

In [33] depth sensor Microsoft Kinect and stereo cameras are combined to obtain high quality depth images. To compensate the low-resolution of Microsoft Kinect devices, nearest-neighbor up-sampling is used. The two depth maps are combined using a energy minimization approach exploiting alpha expansion and Graph Cuts. Results are good, also because the algorithm runs on high quality $4 - 12$ Megapixel input images. However, it is highly unlikely that real-time execution can be achieved with this method, since the actual optimized implementation took over 20 minutes to produce the fused depth map, mainly due to the burden of alpha expansion phase.

Almost all of these approaches do not fully exploit the confidence information of the two measures, furthermore, many methods require to manually assign a weight to the two disparity maps. Fusion methods that rely on global optimization produce good results but are slow, on the other hand, local approaches are usually faster but results are noisy and sometimes the overall improvement is negligible. The best trade-off would be a technique that locally performs a sort of global reasoning restricted to the neighbors. These requirements have led authors of [9] to extend the Locally Consistent (LC) technique [19] originally proposed for stereo matching, to deal with the two disparity hypothesis. In this work, the contributions of the stereo and the ToF acquired data are simply averages, what it is still missing is the definition and association of weights to these two disparity maps for a reliable fusion. In this thesis the same algorithm has been exploited and improved to fuse different hypothesis according to their plausibility.

Chapter 2

Set-up

2.1 Acquisition system

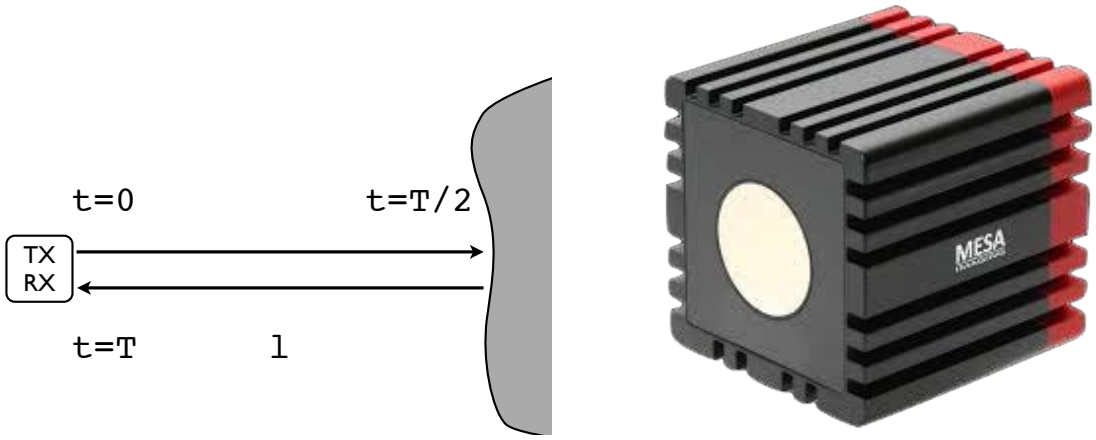
Before describing the details of the devised fusion algorithm, a general introduction to the two family of sensors used in this framework is necessary. Regular cameras are nowadays a well known instrument, Time-of-Flight range sensors and stereo vision systems instead are quite uncommon in the daily life. In this chapter, after a brief introduction to these two families of sensors, the overall acquisition system is presented, as well as some basic terminology and practical assumptions used in this thesis.

2.1.1 Basic of Matricial Time-of-Flight cameras

The distance measurement capability of this family of sensors is based on the Time of Flight principle, that consists in illuminate an object with a light and analyze the reflected ray. In this thesis, the word ToF will always refer to Time-of-Flight cameras, i.e. cameras based on this technology.

The working principle is very simple, as can be seen in Figure 2.1a. Distance l between the sensor and the target can be estimated with just the knowledge of the time T that an electromagnetic wave takes to travel that distance.

Since the electro-magnetic radiation travels in the air at the constant light speed $c \approx 3 \times 10^8 \,\mathrm{m\,s^{-1}}$, the distance covered by an optical radiation in a certain time is simply the product between light speed and time. The required Hardware consists of just a radiation emitter (TX) and a receiver (RX), ideally both co-positioned. At time $t = 0$ the transmitter emits a light pulse that travels straight toward the scene for a distance l, reaching a target point at time $t = T/2$. It is then reflected back, and at time $t = T$ it is received by RX. The relationship between time and distance in this case is



(a) LIDAR **(b)** Matricial ToF [21]

Figure 2.1: Time-of-Flight working principle

$$l = c\,\frac{T}{2}. \tag{2.1}$$

This simple relation allows to measure the distance of a single point. This sensing technology is also called LIDAR (from *light* and *radar*) and it is widely used in different fields to make high resolution depth maps. It requires point-by-point ToF measurements, therefore this simple sensor is intrinsically not suited to acquire dynamic scenes.

Matricial Time-of-Flight cameras are an extension of LIDAR sensors, in which the entire scene is captured by a matrix of $N \times M$ ToF sensors (Figure 2.1b) at one shot, allowing the delivery of depth maps at video rates, which is a fundamental requirement for real-time applications. Due to physical limits, it is not possible to have a one-to-one correspondence between emitters and receivers, however IR LEDs can be positioned in a regular configuration to simulate a single emitter at the center of the receiver matrix. Current technology allows the integration of $N \times M$ receivers on a single CMOS chip, making these sensors very compact and easy to handle, without moving parts. Moreover, the depth map and other useful information are directly obtained without additional computation by the user. Some drawbacks of this technology are the relatively low resolution with respect to a regular camera, poor quality along the region with depth discontinuity and high sensibility to illumination changes, due to other light sources in the scene like the sun, or to the lack of reflectivity in the acquired objects.

Additional information on this technology in Chapter 3 is discussed, and a detailed review of the state-of-the-art in ToF technology in [34] can be found.

In the following discussion, often the reference Time-of-Flight camera will be the MESA SR4000, since this is the model used to validate the analysis.

2.1.2 Basic of stereo vision

A stereo vision system, called simply stereo from now on, is a framework made by two regular cameras (typically identical), that exploits the same human stereopsis paradigm to provide an estimate of depth distribution of the scene framed by the two cameras.

Stereopsis, also known as binocular vision, is the process that allows our brain to extract information on the tridimensional structure from a pair of slightly different images of the same scene captured by the two eyes. The same concept can be applied to a pair of cameras framing the same scene, separated by a certain distance, exactly like our eyes. It is common to use the left camera (denoted by L) as reference viewpoint: from this assumption, the other common naming *reference* for the left camera and *target* for the right one (denoted by R) follows straightforwardly.

The 3D position of a point can be inferred by means of triangulation of correspondent points. Starting from a simplified case in which the two cameras are parallel and aligned (Figure 2.2), also called standard form, consider a point $\mathbf{P} = [x, y, z]$ in the space and the projections $\mathbf{p}_L = [u_L, v_L]$ and $\mathbf{p}_R = [u_R, v_R]$ in the two camera image planes, left and right respectively. Triangulation is the process of determining the coordinates of $\mathbf{P}$, especially the depth coordinate z, given its projections $\mathbf{p}_L$ and $\mathbf{p}_R$.

In this simple case where cameras are rectified, it is easy to understand that the only difference in the coordinates of $\mathbf{p}_L$ and $\mathbf{p}_R$ is in the horizontal coordinate u, as the vertical coordinate v will be the same. Given the geometry depicted and similar triangles properties, the following equations can be derived

$$\begin{cases} \dfrac{f}{z} = \dfrac{-u_L}{x} \\ \dfrac{f}{z} = \dfrac{-u_R}{x-b} \end{cases} \tag{2.2}$$

from which after some manipulation we obtain

$$z = \frac{b\ f}{u_R - u_L} = \frac{b\ f}{d} \tag{2.3}$$

In the previous equations, f is the focal length of the two cameras, b is the distance between the two optical centers, also known as *baseline* and $d = u_R - u_L$ is the so

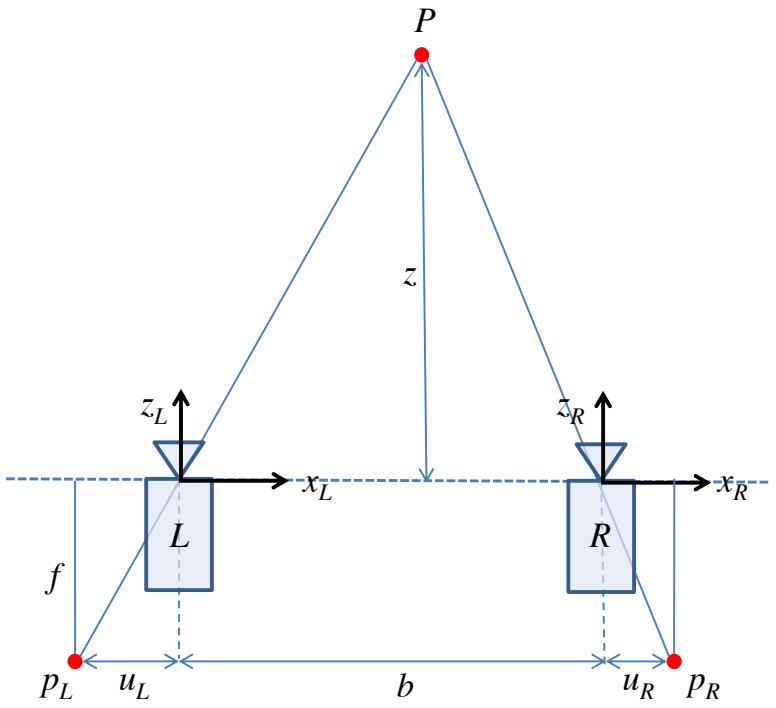


Figure 2.2: Stereoscopic triangulation (image from [8])

called *disparity* associated to point $\mathbf{p}_L$, i.e. the difference between x coordinate of the two corresponding points in left and right image planes. Equation (2.3) shows how it is possible to retrieve the third component z when disparity and geometry of the system are known.

While f and b can be estimated with a procedure called *calibration*, the disparity d requires to find corresponding points, also known as *conjugate points*, in the two images. Given a point $\mathbf{p}_L$ in the left image, the correspondent point $\mathbf{p}_R$ in slave image has to be found. We know that the two images are not so different, however the correspondent point could be at any pixel. A search of that point in the entire image would require a lot of complexity, also because the most common similarity criterions require to do operations in a window for every pixel. Fortunately, the search domain can be limited to a one dimension (along u) thanks to epipolar constraint. A geometrical analysis shows that the conjugate point of $\mathbf{p}_L$ in the second image, must lie in a straight line called epipolar line of $\mathbf{p}_L$.

In a more realistic scenario the two cameras are not perfectly aligned, however it is always possible to apply a linear transformation to images acquired by the two cameras in order to simplify the task of correspondence selection. This procedure called *rectification* is briefly discussed next in this chapter.

The fundament theory behind stereo vision can be found in classical computer

vision books such as [35, 16]. In Chapter 4 more details about stereo algorithms and disparity computation are discussed.

2.1.3 Overall imaging system

The acquisition framework consists in a stereo vision system (denoted by S) made by two regular cameras, and a Time-of-Flight camera (called T). Although the three devices do not require a specific arrangement, it is customary to place T between the two cameras. Furthermore, the two imaging systems have to acquire the same scene, therefore T and S have to be as close as possible. Figure 2.3 shows the actual arrangement of the overall system.

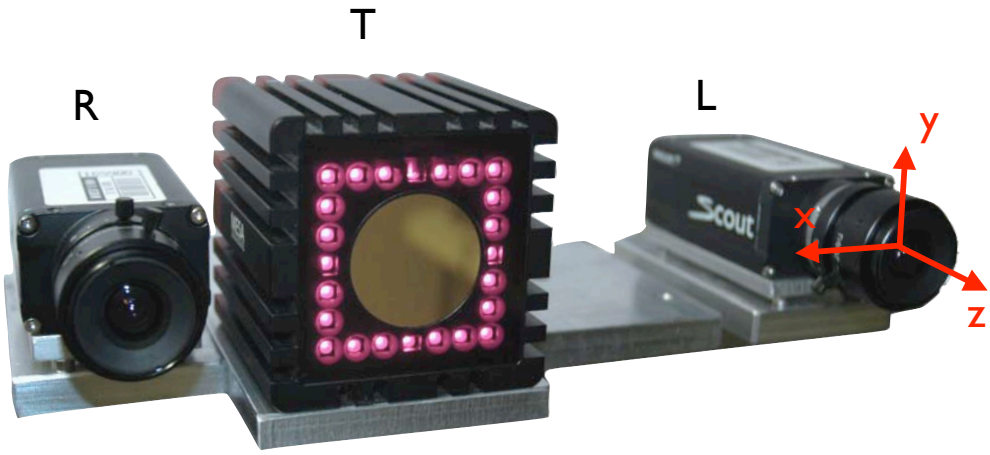


Figure 2.3: Acquisition system: ToF and stereo

Data fusion requires to have maps referred to the same viewpoint: in this thesis the left camera has been used as reference system. The reference 3D camera coordinate system (CCS) is colored red in Figure 2.3.

In this work only static scenes have been considered; however, in case of dynamic acquisition, it is also useful to have a synchronization unit tool, providing frames at the same time.

The set-up made by multiple measurement instruments of different nature, considering the properties of the same object, is usually called *heterogeneous measurement system*. In this framework, visual quantities like color and geometry of the system have to be acquired, which are obviously highly dependent on the spatial position of each instrument. Consequently, the measurements can only be related upon the knowledge of the internal characteristics of each sensor and their relative positions. This information can be retrieved by means of the *calibration* procedure, i.e. the estimation of the relationship concerning measured

and actual correspondences between camera sensor pixels and 3D scene points. For heterogeneous systems, calibration requires also to estimate the relative position of the multiple sensors.

Calibration of stereo system has been widely studied and standard procedures are now available [3]. ToF cameras image formation can be modeled by perspective projection, therefore ToF camera calibration is similar to the procedure for a regular camera. The joint calibration of the two systems for data fusion requires even more attention: not only the intrinsic parameters of the two systems have to be estimated, but also their relative position must be determined with high precision. In [7] a generalized multi-camera calibration technique for ToF and stereo cameras is described and in [8] an exhaustive review of both generic acquisition systems and heterogeneous systems can be found.

Chapter 3

Matricial Time-of-Flight Range Camera

In Chapter 2, the Time-of-Flight working principle has been introduced. Despite its conceptual simplicity, the actual implementation requires great effort to all of the major manufacturers of ToF cameras like MESA Imaging [21], PMD Technologies [28], SoftKinetic [32] and Microsoft [22], mainly because measurements require high precision at a clock period of few pico seconds. From 2.1, for example, it can be seen that a resolution of 1 mm requires at least a clock period of 6 ps, i.e. the time necessary to light pulse to travel back and forth that distance. However, the accuracy of the depth measurements is subject to errors due to many other factors, which can be categorized either internal, due to noise or calibration, or environmental, if they can be attributed to effects dependent of the scene viewed.

Different approaches have led to different technologies, although the most common adopted by commercial solution is the continuous wave (CW) intensity modulation. Information on other techniques such as optical shutter (OS) and single-photon avalanche diodes (SPAD) can be found in [26, 18].

This chapter provides a general introduction on CW ToF cameras and on related practical issues, with a description of performance limits and noise characterization. This information is needed to later understand the confidence maps built on top of these issues. Depth data acquired by the ToF cameras need to be up-sampled to the spatial resolution of the stereo vision images, so a novel up-sampling algorithm based on image segmentation and bilateral filtering is also described.

3.1 Operation principles

As an electromagnetic signal can be propagated if modulated by a sinusoid of a certain frequency f_{mod}, a modulated infrared (IR) wave of amplitude A_e is sent from the emitter toward the target. After reflection from the scene, the sensor receives back a signal with a mean offset I, accounting for background illumination and camera electronics, and of amplitude A. The phase delay between these two signals[1] is proportional to the distance of the observed point (Figure 3.1).

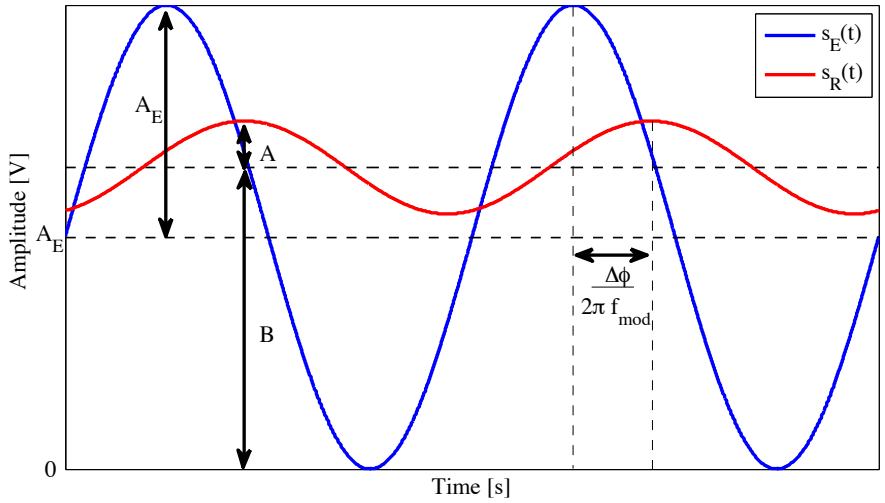


Figure 3.1: Example of modulated signals for ToF measurement (image from [8])

The received signal can be demodulated using cross correlation with respect to the reference sinusoid. The resulting signal can be written as

$$s_R(t) = A\cos(2\pi f_{mod}t + \varphi) + I \tag{3.1}$$

If we consider four samples s_R^i of this signal at four phase intervals, for instance at $t_i = k/4f_{mod}$, for $k = 0, \ldots, 3$, amplitude A, intensity I and phase φ of the received signal are given by [24]

[1] In IR based ToF camera, due to finite bandwidth of the IR-LED, it is assumed that only the fundamental harmonic of the modulated frequency is transmitted.

$$
\begin{aligned}
A &= \frac{\sqrt{\left(s_R^2 - s_R^0\right)^2 + \left(s_R^3 - s_R^1\right)^2}}{2} \\
I &= \frac{s_R^0 + s_R^1 + s_R^2 + s_R^3}{4} \\
\varphi &= \arctan\left(\frac{s_R^3 - s_R^1}{s_R^0 - s_R^2}\right)
\end{aligned}
\tag{3.2}
$$

Amplitude and intensity, as we will see later, are useful for SNR computation, while, given φ, the distance l can be obtained as

$$
l = \frac{c}{4\pi f_{mod}}\,\varphi \tag{3.3}
$$

Some of this information can be retrieved from the data provided at frame rate from the camera for each pixel. MESA SR4000 for example provides:

- *Depth map* with values quantized to 14 bit and expressed in meters from 0 to $l_{MAX} = c/2f_{mod}$;
- *Amplitude map* that contains the estimated amplitude of the received signal quantized to 16 bit and scaled to be independent of distance in the image array;
- *Confidence map* that represents a measure of probability of how correct the distance measurement is expected to be. Values are 16 bit quantized and the estimation involves distance and amplitude of measurements as well as their temporal variation.

3.2 Practical issues

The ideal derivation previously discussed actually involves a number of practical implementation issues that must be taken into account. The main non-idealities that have been studied in literature are described in the following list.

Phase wrapping From Equation (3.2), the phase delay is obtained from an arctangent function. The original $[-\pi/2, \pi/2]$ codomain interval can be extended, with the usage of $\arctan 2(\cdot,\cdot)$ function, to $[0, 2\pi]$. The substitution of these limits into Equation (3.3) shows that the estimated distance can be in the range $[0, c/2f_{mod}]$. More generally, the distance corresponds to the reminder

of the division between ϕ and 2π, multiplied by $c/2f_{mod}$, as ϕ is estimated modulo 2π. To overcome this problem, the usage of non-sinusoidal wave-form has been tried.

Harmonic distortion As one method to generate sinusoids is to filter a squared wave-form with a low-pass. and the sampling of the received signal is not ideal, these two inaccuracies lead to a distortion in the estimated phase and consequently in the estimated distance. This is a systematic offset that can be reduced by means of a look-up-table correction.

Different reflectivity The amount of reflected light strongly depends on the reflectivity of the target object, which leads to erroneous distance calculation. Materials can be divided into two categories according to their reflection coefficient in the IR band of the emitters. For *diffusely reflecting materials* such as dull surfaces, the reflectivity coefficient has values in the range $[0, 1]$, where 0 means that all incoming light is absorbed or transmitted, and 1 that all the incident rays are reflected. The reference value of 1 is given by the case of a perfect Lambertian reflector, where all the light is back-scattered with an intensity distribution that is independent of the observation angle. For *directed reflecting materials* such as glossy surfaces, the reflection coefficient might be even ≥ 1 for specific angles at which the light is directly reflected into the sensor. Camera measurements for such directed reflections might saturate, causing errors in distance estimation. The same problem may be encountered in the opposite condition, that is when the reflected ray points away from the camera, preventing the sensor from capturing enough signal intensity to deliver valid measurements. Authors of [37] proposed a method to correct the distance non linearities as well as the integration time offsets for different reflectivity. They found that a difference in amplitude as well as measured distance between the black and white targets are attribuited to the differences in reflectivity. In [11] it is shown that the systematic error in depth measurement can be reduced using the object's intensity. Depth and inverse amplitude $1/A$ are compared, discovering that these two measures are correlated.

Angle incident Quality of the received signals also depends on the angle at which the light is emitted, reflected or received. In [37], the model to correct distance nonlinearity also considers a term related to the angle of emitted and received rays. Moreover, materials with different reflection coefficient impact the measurement characteristics of the camera in different ways. The

best measure is given by the case of Lambertian reflection of a 90° incident and received ray. As a prior knowledge about objects material composition and orientation in the scene is not available, modeling this inaccuracy is a quite difficult task. The only information that is always known is the angle associated to the emitted light rays. A general characterization of this phenomenon is available in the datasheet of the actual camera. MESA SR4000, for example, defines two measurement regions (Figure 3.2): the first region involves central pixels while the second one involves pixels far away from the center point. A larger error is associated to the outer region, and this is due to the larger angle of the emitted light rays. This indication of the measurement accuracy is also known as *repeatability* and is characterized by the spread σ of the measurement around the mean value.

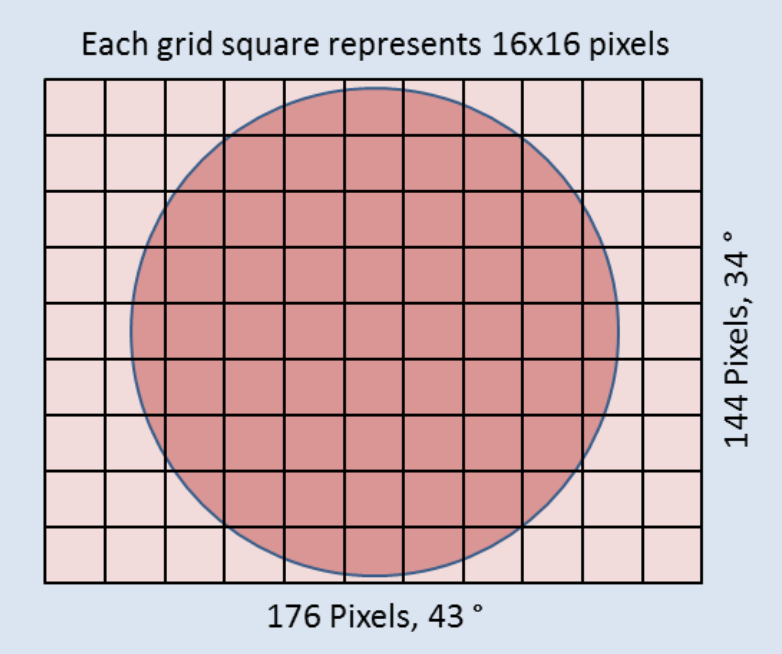


Figure 3.2: ToF measurement regions with different repeatability

Photon shot noise ToF cameras suffer from various noise sources common to all light-collecting sensors. Thermal noise, reset noise, flicker noise, dark current noise and quantization noise can be mitigated by lowering the operating temperature or improving fabrication techniques. The principal noise source that cannot be suppressed by a preliminary calibration is the sthocastic photon shot noise. The samples of the received signal s_R^i are obtained as the sum over the actual number of photons collected by the receiver over a finite period of time, so they can be caracterized by a Poisson process. Accuracy and performance for a ToF camera can be measured by the quality of the

amplitude A with respect to the underlying noise. From the detailed analysis in [24], it turns out that the signal-to-noise ratio of $\hat{A} \sim \text{Rice}(A, \sqrt{I/2})$, where $\hat{A}$ represents the actual measured amplitude value, is a suitable metric for this measure. For the received signal with Rice distribution, the definition of SNR is

$$SNR = \frac{A}{\sqrt{I/2}} \tag{3.4}$$

that can be thought as the ratio between the signal amplitude and the standard deviation of the additive Gaussian noise corrupting the received signal. A more realistic estimation for low SNR cases, exploiting a maximum likelihood approach, is provided in the same article, too.

Saturation and background light Most of the internal noise sources can be limited by averaging the measured values over several periods, and better repeatability of distance measurements is achieved with long integration time[2] without reaching saturation. In photography, the analogous parameter is the exposure time, a method used to control the amount of light recorded by the camera's sensor. Long exposure times allow to obtain brighter images, but increase the likelihood of saturation. This phenomenon is particularly critic in presence of external IR illumination or in case of highly reflective objects. In general, a non negligible background light decreases the signal to noise ratio of the received wave. An optimal trade-off in the choice of integration time is necessary to obtain the best measurement precision without saturation. Algorithms for automatic integration time setting and background light suppression have been developed for example in [4, 5].

Motion blur Long integration time may be a solution in the attempt of noise reduction. However, long times might introduce elements of blur in the image, in addition to saturation issues as previously described. These effects are even more visible where a dynamic scene has to be acquired. Algorithms for optimal integration time setting must take into account also the maximum level of blurring allowed without compromising the measurements quality.

Finite size area A sensor pixel is ideally associated to a point of the scene, but actually this region has a non negligible area. If this region entirely belongs to the same object (same material and color) and all the points inside this

[2]Integration time is the length of the averaging time.

region are at the same distance from the camera, the approximation that a single point is associated to a pixel is valid. Otherwise, especially if the area crosses a depth discontinuity, the resulting depth estimate is an unpredictable value, and the pixel associated to such values are called *flying pixels*.

Multipath The depth of each point in the scene is estimated using (2.1), therefore the distance traveled by the light is about twice the distance from the camera and the object. However, due to multiple reflections, the light may reach the object or the receiver along several paths and therefore the measured distance may be over-estimated with respect to the true distance. This effect is easily visible when measuring objects that have concave structure, such as corners between two walls. Given its scene-dependent nature, currently there are no valid methods to compensate for this inaccuracy, and multipath effect is very hard to model.

Multiple camera ToF camera projects IR light in the scene and, if multiple cameras are running at the same time, the emitted rays may disturb each others' measurements. There exist different possibilities to overcome this problem, for example exploiting time multiplexing, in which a control system enables the measurement of individual cameras consecutively, or frequency multiplexing, in which the light is collected in the other systems only as background illumination without perturbing the distance measurement.

3.3 High resolution disparity map from ToF data

Depth map provided by a ToF camera has a resolution much lower than that of a regular camera. However, the fusion framework requires to deal with data of the same resolution, therefore an up-sampled depht map for this kind of sensor is needed. In this thesis a novel algorithm introduced in [9] is exploited to interpolate the sparse depth map, using both segmentation and bilateral filtering. This allows to combine the good edge preserving quality of the segmentation-based methods and the good robustness of the bilateral filter.

Starting from the low resolution depth map provided as output by ToF camera, the procedure to obtain the up-sampled depth map is now described.

Undistortion and rectification In order to revert the effects of lens distortion, depth image has been undistorted; furthermore, since the stereo imaging system required to be rectified, also ToF data images have been rectified.

Projection to reference Depth measurement acquired by the ToF camera need to be projected into the image plane of the reference camera to have the same viewpoint. Each point $\mathbf{p}_T$ of the low resolution ToF lattice is first back-projected using the intrinsic parameters matrix K_T and the z coordinate represented by the pixel value, obtaining the 3D point $\mathbf{P}_T$. Then a roto-translation from ToF reference system to the left camera reference system is applied, obtaining a 3D point $\mathbf{P}_T^M$ referred to left camera. A final projection using left camera intrinsic parameters matrix K_M provides pixel coordinates in the reference camera lattice $\mathbf{p}_T^M$. The overall transformation produces a ToF sparse depth map seen from left camera perspective. The support of this map is just a small subset of the camera lattice (usually the number of assigned pixels is below 10%).

Occlusions removal The new depth map obtained from the previous step needs to be refined. ToF and left camera have different points of view: as a consequence the scene seen by the two cameras is slightly different and some of the points originally in the ToF field of view may no longer be visible after the projection. These occluded points have been removed by means of an hybrid technique between scanline rendering and z-buffer. When two or more 3D points are associated to the same pixel, only the point closest to the camera is actually visible. Rather than pixel-by-pixel basis, this algorithm works on a row-by-row basis, where triangular primitive and a depth buffer are used to determine the closest points.

Segmentation If the knowledge of depth discontinuity would be available, a smart interpolation technique could be applied to the sparse point cloud, to obtain a high resolution depth map. However, low resolution depth data also results in poor edges localization, therefore the color image of reference camera is suited to improve the spatial resolution. Segmentation is a process that divides the color image in a set of regions, called segments, ideally corresponding to the different scene objects. It is reasonable to assume that inside a segment the depth varies smoothly and that sharp depth transitions between different scene objects occur at the boundaries between different segments. The algorithm implemented is based on mean-shift clustering proposed in [6]. Although the large research activity in this field, currently there are no procedures completely reliable for any scene, hence segmentation artifacts can lead to errors in the next step.

Interpolation The goal of this final step is to associate to all the points of the

camera lattice a depth value. In order to accomplish this, a window W_j of size $w \times w$ centered on each of the p_j samples that does not have a depth value already available is considered for the computation of the estimated depth value $\tilde{z}_j$. The samples that already have a disparity value from the ToF measures will instead just take that value. The set of points inside the window can be denoted with $p_{j,k}, k = 1, ..., w^2$ and finally $W'_j \subset W_j$ is the set of the points $p_{i,k} \in W_j$ with an associated depth value z_i. In standard bilateral filtering [36], the interpolated depth of point p_j is computed as the weighted average of the depth values in W'_j, where the weights are computed by exploiting both a weighting function in the spatial domain and one in the range domain. In the cross bilateral filtering, a standard 2D Gaussian function as in [36] is employed for the spatial domain weighting function $f_s(p_{i,k}, p_j)$, the range domain function $f_c(p_{i,k}, p_j)$ is also a Gaussian function but it is not computed on the depth itself, but instead on the color difference in the CIELab space between the two samples. In order to exploit the segmentation information to improve the performance of the bilateral filter, authors of [9] added an additional third indicator function $I_{segm}(p_{i,k}, p_j)$ defined as

$$I_{segm}(p_{i,k}, p_j) = \begin{cases} 1 & \text{if } S(p_{i,k}) = S(p_j) \\ 0 & \text{if } S(p_{i,k}) \neq S(p_j) \end{cases} \tag{3.5}$$

The interpolated depth values are finally computed as:

$$\begin{aligned} \tilde{z}_s^j = \sum_{W'_j} \big[& f_s(p_{i,k}, p_j) I_{segm}(p_{i,k}, p_j) z_{i,k} + \\ & f_s(p_{i,k}, p_j) f_c(p_{i,k}, p_j)(1 - I_{segm}(p_{i,k}, p_j)) z_{i,k} \big] \end{aligned} \tag{3.6}$$

This interpolation scheme acts as a standard low-pass interpolation filter inside each segmented region. Samples that are outside the region, instead, are weighted on the basis of both the spatial and range weighting functions thus getting a lower weight. The inclusion of samples outside the segmented region ensure the robustness with respect to segmentation artifacts. Performance of this method is limited by two main issues, namely by segmentation errors and by inaccuracies due to depth acquisition or to the calibration between ToF and color cameras.

Once the high resolution depth map is obtained, by relation (2.3) the correspon-

dent disparity map can be computed. It is worth repeating that for Time-of-Flight cameras the corresponding disparity map has no real meaning, as this is a concept related to binocular vision, but the conversion remains valid.

3.4 Confidence estimation of ToF disparity

Practical issues described in Chapter 3.2, allow the derivation of different confidence maps. Due to the quite recent introduction of ToF cameras, most of the applications that use these sensors rely on manufacturer's calibration and use directly the confidence map provided by the ToF camera to discard those points with a bad confidence. Other works instead rely on amplitude values as an indicator of confidence, however, [30] and others show that simply thresholding low-amplitude values is insufficient to remove inaccurate pixels. These and other more reliable confidence maps for ToF disparity map are now described.

Since they are associated to the up-sampled disparity map, also confidence maps have to be of the same high resolution.

3.4.1 Confidence from ToF amplitude map

One of the data matrices provided by a ToF camera is the amplitude map. Since all the measurements are extracted from samples of the received signal, its amplitude level can be considered a good measure of confidence. In contrast to the confidence image, amplitude map is directly provided by the camera and no extra processing resources are needed.

The received amplitude highly depends on the reflection characteristics of the scene and objects that are acquired. As described in the previous section, the reflectivity of the target object has a large influence on the repeatability of the measurements. Furthermore, measured distance also affects the received wave's amplitude. Signals coming from further points will have a lower amplitude, due to the attenuation parameter in the electromagnetic wave propagating through a non ideal medium. This factor is proportional to the square of the measured distance. A separate confidence measure directly exploiting the distance of objects in the scene could be considered, however the main effect of distance is related to the amplitude, and this map already takes into account the depth information.

This information cannot be directly exploited to provide a likelihood for each point in the high resolution disparity map, because of the difference in their resolutions. The procedure to find the correspondent pixel in the original amplitude image, given the high resolution depth image, requires to go over the same procedure

of the *projection to reference* step to compute the high resolution depth map, but in the opposite order. Starting from the up-sampled depth image computed before, each pixel in the image $\mathbf{p}_L$ is first back-projected to the 3D world using the intrinsic parameters matrix of the left camera K_L and the pixel value as z coordinate, obtaining the point $\mathbf{P}_L$. Then a roto-translation from left camera reference system to ToF reference system is applied, obtaining a 3D point $\mathbf{P}_L^T$ in the ToF viewpoint. A projection using intrinsic parameters matrix K_T, provides the pixel coordinates in the ToF lattice $\mathbf{p}_L^T$. The amplitude value stored at that pixel in the undistorted amplitude map is finally associated to the original pixel $\mathbf{p}_L$ of the confidence map under construction. As a single pixel in the low resolution lattice corresponds to multiple pixels in the high resolution lattice, there will be confidence values associated to the same amplitude value. This process requires the high resolution depth map to be available, therefore it is not possible to project amplitude values together with depth measures in the same step, otherwise some pixels in the high resolution amplitude map will not have a valid value.

When confidence maps were introduced in Chapter 1, a requirement was to have values in the range $[0, 1]$: a normalization to that interval is therefore needed.

3.4.2 Confidence from ToF confidence map

Some ToF camera provides as output also a confidence measure of the estimated depth and it is quite reasonable to directly use this matrix as measure of likelihood. MESA SR4000, for example, has a confidence map and its manual states that low confidence is typically associated to low reflected signal or movement in the scene. In addition, a visual comparison of the confidence map in different scenarios, with the measurement regions in Figure 3.2, highlights the correlation between these two data. It is likely to assume therefore that the algorithm for confidence estimation forces the values to be lower in the outer region than the central region, decreasing with radial distance from the center.

The generation of confidence map is delegated to the driver of the PC connected to the camera, using a combination of distance and amplitude measurements and their temporal variations. An extra processing is therefore necessary, however it is efficiently implemented and the resulting computation time and complexity are negligible.

This confidence measure takes into account different practical issues of ToF cameras described in Chapter 3.2: amplitude effects have been already discussed in the previous section, distance is also explicitly considered in this measure since further points exhibits lower confidence, and finally angle of incidence of the

incoming rays, given that central pixels being more accurate present a higher confidence.

The procedure to obtain the high resolution confidence map requires to go over the same operations described for the previous confidence estimation. In this case, in the last step the undistorted confidence map is used instead of amplitude map. Also in this case the resulting values have been normalized to the unit interval $[0, 1]$.

3.4.3 Confidence from Amplitude and Intensity

As presented in Chapter 3.1, the SNR of the received signal can be approximated by the ratio

$$SNR = \frac{A}{\sqrt{I/2}} \tag{3.7}$$

and according to [24], [4] and [5], the probability density function of the noise affecting depth estimate can be approximated by a Gaussian with standard deviation

$$\sigma_z = \frac{c}{4\pi f_{mod}} \frac{1}{SNR} = \frac{c}{4\pi f_{mod}} \frac{\sqrt{I/2}}{A} \tag{3.8}$$

The standard deviation σ_z determines the precision of the range measurement.

This formula proves from another point of view that when amplitude A increases, precision improves as the standard deviation decreases. The same equation suggests also that as the interference intensity I increases, the precision gets worse. Intensity I 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 I in (3.8). Finally it is worth to observe that if modulation frequency f_{mod} increases the precision improves. Modulation frequency is also related to phase wrapping and so to maximum measurable distance. If f_{mod} increase measurement accuracy improves but maximum measurable distance decrease and vice-versa. Therefore longer distance can be measured at the cost of decreasing the precision.

Signal to noise ratio in Equation (3.7) or depth standard deviation (3.8) could be directly used as confidence measure. However, the fusion algorithm is based on disparity map and better performance could be achieved if a confidence based on disparity rather than depth can be exploited. For a given distance z, if a certain depth interval Δ_z around z is considered, the corresponding disparity interval Δ_d

also depends on the distance z [40], due to the inverse proportionality (2.3) between depth and disparity. The goal is thus to find the corresponding standard deviation of the disparity measurement σ_d.

Since the depth noise has been approximated by a Gaussian, if a certain point is estimated at depth μ_z, this can be interpreted as a Gaussian with mean μ_z and standard deviation σ_z. As depicted in Figure 3.3, a depth interval $\Delta_z = |z_1 - z_2|$ around the mean value can be considered, for example $[\mu_z - \sigma_z, \mu_z + \sigma_z]$ leads to $\Delta_z = 2\sigma_z$.

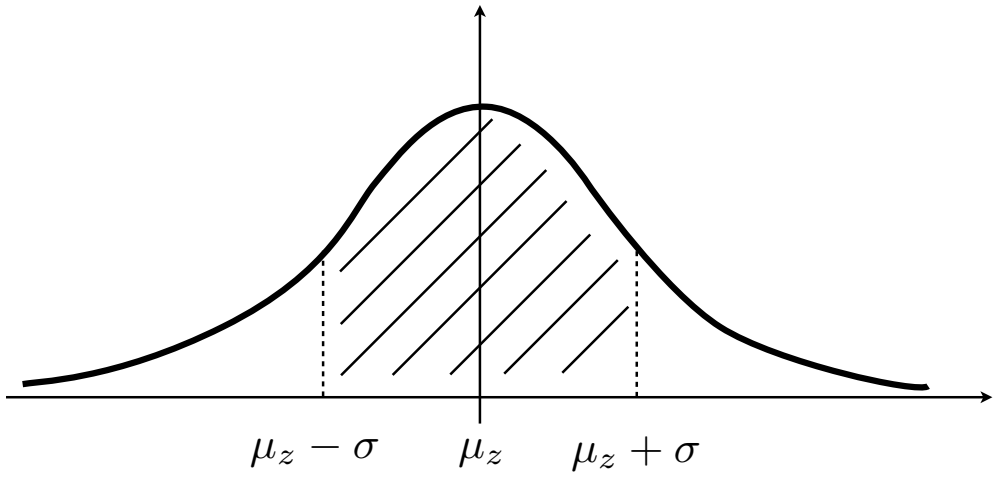


Figure 3.3: Gaussian of depth measurement

In order to find the corresponding interval Δ_d in the disparity measure, the following can be observed

$$\Delta_d = |d_1 - d_2| = \frac{bf}{\mu_z - \sigma_z} - \frac{bf}{\mu_z + \sigma_z} = bf\ \frac{2\sigma_z}{\mu_z^2 - \sigma_z^2} = 2\sigma_d \tag{3.9}$$

where d_1 and d_2 are the correspondences of z_1 and z_2. From the last equality it follows that

$$\sigma_d = bf\ \frac{\sigma_z}{\mu_z^2 - \sigma_z^2} \tag{3.10}$$

Equation (3.10) provides thus the corresponding standard deviation of the noise in the disparity measure. This value is also affected by the mean value of the measurement, unlike the standard deviation of the depth measurement, and this is consistent with the inverse proportionality of depth and disparity.

Amplitude and intensity values have to be selected with the down-sampling

look-up procedure described before. If the intensity map I is not available from the ToF camera, the grayscale image from left camera can be used.

The relation between standard deviation and confidence in Figure 3.4 is depicted.

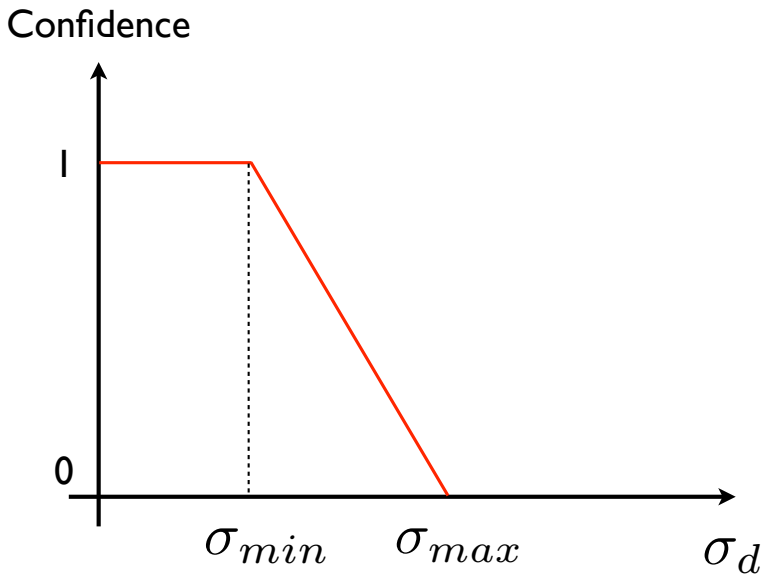


Figure 3.4: Relation between disparity standard deviation and confidence

Theoretically, standard deviation could assume values in a big interval and in order to guarantee confidence values to be plausible, two thresholds (σ_{min} and σ_{max}) have to be fixed. These two values are not critical to select and have been introduced just to guarantee a certain limited bound.

The usage of σ_d instead of σ_z allows to better weight the confidence in the disparity map. A small error in depth measurement for a close object will result in a bigger error of disparity measure and consequently in a confidence reduction.

The confidence map built from ToF amplitude map can be considered as a special case of this measure. When the illumination I is constant, indeed, the two measures just differ from a constant. Since this likelihood measure takes into account more parameters, better results are expected.

3.4.4 Confidence from local variance

One of the limitations of confidence models described so far is that they do not take into account the practical issues of finite size sensor pixels. In order to account for such non ideality, another confidence map is proposed.

When the scene area associated to a pixel comprises two regions at a difference depth, the resulting estimated range is some convex combination of the two depth. This effect is presumably associated to all the discontinuity, and given that it is not

known a priori which value is selected, it is reasonable to associate to these regions a low likelihood. It is worth exploiting the fact that if pixel p_i is associated to a scene area crossed by a discontinuity, some of the points p_j in the 8-neighborhood $\mathcal{N}(p_i)$ of p_i are relative to points at a closer distance, and some others to points at further distance. 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 mean deviation of the points in $\mathcal{N}(p_i)$ with respect to p_i will be higher than if the 8-neighborhood belongs entirely to the same depth region.

For each pixel p_i the local variance σ_l^2 can be computed as

$$\sigma_l^2 = \frac{1}{|\mathcal{N}(p_i)|} \sum_{j \in \mathcal{N}(p_i)} (z_i - z_j)^2 \tag{3.11}$$

where $|\mathcal{N}(p_i)|$ is the cardinality of the neighborhood considered, in this case equal to 8, and z_i and z_j are the depth values associated to pixel p_i and p_j respectively.

A variant to this formula could be to associate a lower weight to further points, however, assigning the same weight to all the pixels in the 8-neighborhood, allows to detect more discontinuity, both straight edges and angular edges.

This computation is performed for every pixel with a valid depth value. It may happens however that some p_j considered in a 8-connected patch do not have a valid value. This is usually due to occlusions that imply with high probability a depth discontinuity. In order to obtain a reliable map, a constant value can be used in the summation (3.11) in place of $(z_i - z_j)^2$ for those occluded pixels p_j.

The resulting image, after a normalization to unit interval and a simple interval remapping, provides a valid confidence map. Points where the local variance is high are associated to discontinuity, therefore a low confidence should be assigned. Where instead the local variance is close to zero, a higher confidence should correspond. By calling $\bar{\sigma}_l^2$ the normalized local variance, the formula that relates confidence C and $\bar{\sigma}_l^2$ is simply

$$C = 1 - \bar{\sigma}_l^2 \tag{3.12}$$

An improvement of this confidence map could be to consider a variable window size according to the depth. Closer edges are associated to a bigger area, therefore a bigger window may improve the confidence estimation.

3.4.5 Overall confidence

From the four confidence maps previously described, one global map can be derived according to these observations:

- Amplitude of the received signal is a valid measure of the range measure goodness, however high values are not always associated to good measures. It may happens that when the amplitude is high, also the background illumination is high, causing a worsening in quality measure. Similarly, when the amplitude is low, if also the intensity is low, the measure should not be considered of poor quality.

- Confidence provided by ToF camera takes into account different parameters such as distance and amplitude. This information suffers therefore of the same problems affecting confidence from amplitude.

- Range accuracy analysis improves the lack of confidence from amplitude. It provides a robust model against false low or high amplitude points, accounting also for received intensity. The detailed comparison in Chapter 6 will prove the reliability of this analysis. This map however does not take into account the effects of the finite size area associated to a pixel.

- What is missing in the range accuracy analysis is the characterization of the measure along edges, therefore another confidence measure has been introduced. The uncertainty derived from a non-deterministic range association along regions with depth discontinuity has led to associate to those regions a low confidence. It is not always true, but with high probability depth measurements along edges will be little accurate.

From these observations it can be deduced that amplitude and confidence maps provided by ToF camera, considered individually, do not have enough information to well describe all the practical issues of real ToF cameras. A direct combination of these two images like a simple pixel-by-pixel product still would not be sufficient to produce an accurate confidence measure. On the other hand, range accuracy analysis involves more information and it is able to handle critical cases. In addition, confidence from local variance provides the necessary information to assign to edges and depth discontinuities low confidence.

The global confidence map that has been considered is the pixel-by-pixel product of confidence from range accuracy and confidence from local variance.

As one may expect, the multiplication by local variance confidence just account to lower the likelihood provided by range accuracy analysis along depth discontinuities. Experimental results in Chapter 6 will compare these confidence measures in order to analyze the actual behavior in real scenarios.

Chapter 4

Stereo vision system

Although there are a lot of companies producing stereo cameras, one can easily observe that two single standard cameras can be used as well. Commercial stereo products like Point Grey's stereo camera [29] however, provide a solid framework: users do not have to worry too much about baseline, focal length, calibration and synchronization. Moreover they usually provide Software Development Kit (SDK) including drivers, full software library, Application Programming Interface (API), as well as example programs and source code for a quick integration in the most common programming environments such as C/C++. On the other hand, using two single cameras users have more degrees of freedom on system parameters choice, and many commercial products such as Basler [1] provide valid SDKs as well.

In this chapter, before describing the procedure used to compute disparity map and relative confidence maps for stereo architecture, more details on stereo algorithms are discussed. In particular practical issues of correspondence selection, or disparity computation, are analyzed in order to understand the reasoning behind confidence measures.

4.1 Stereo matching algorithms

The goal of a stereo matching algorithm is to couple pixels in one image with corresponding pixels in the other image exploiting some constraints:

Similarity This is implicit in the correspondence problem, points have to be similar in the two images.

Epipolar geometry The conjugate point lies in a straight line as discussed before.

Smoothness Depth of a smooth surface changes slowly, away from edges.

Uniqueness A point in one image must correspond to only one point in the other image. This assumption is violated if there are transparent objects.

Monotonic order constraint If a point p_1 in one image corresponds to p'_1 in the other, the correspondent of another point p_2 that lies to the right (left) of p_1 must lies to the right (left) of p'_1. This requirement fails if p_2 lies in a particular conical region described by p_1 and the two cameras' optical center.

According to Scharstein and Szelisky [31], in almost all stereo algorithms four building blocks can be defined: matching cost computation, cost aggregation, disparity computation and disparity refinement.

The first step is the *matching cost computation*. The most common pixel-based matching costs include sum of squared differences (SSD), sum of absolute differences (SAD), normalized cross correlation (NCC) and census transform. Sometimes a preprocessing stage like Laplacian of Gaussian or bilateral filtering precedes this phase. Subtraction of the mean value in the window may help to improve the robustness against noise and photometry distortion.

Cost aggregation for local and window-based methods is simply a summation or averaging over a support region and can be either two-dimensional, for the simplest cases, or three-dimensional for better supporting slanted surfaces. Aggregation can be performed using convolution or more efficiently exploiting box-filtering.

For *disparity computation* two main approaches can be found: local methods and global methods. Authors of [31] also describe another class, usually called semi-global methods, i.e. algorithms based on dynamic programming and cooperative algorithms.

Local methods consider only local similarity measures between the region surrounding a pixel and regions of similar shape around all the candidate conjugate points on the other image. The window size can either be fixed or variable, in order to better adapt to each point in the scene. The selected disparity is the one maximizing the similarity measure, a method typically called Winner Takes All (WTA) strategy. As will be presented next in this chapter, local methods are not able to deal with many of the practical issues in stereo vision. The result is usually noisy, because the solution is not regularized. A possible solution to this problem is presented in [15] where, to regularize the solution, the authors propose to smooth the cost volume with a weighted box filter. The well known "edge-fattening effect" in stereo due to aggregation over a support window, can be limited with the usage of the recently proposed guided filter [13], which has a runtime independent of the filter size and preserves edges better than the fast approximation of bilateral filter.

Global methods do not consider each couple of points on its own but estimate all the disparity values at once exploiting global optimization schemes. The general objective is to find a disparity function d that minimizes a global energy made by a term that measures how well the disparity function agrees with the input image pair, and a smoothness term defining the smoothness level of the disparity image by explicitly or implicitly accounting for discontinuities. Global methods based on Bayesian formulations are currently receiving great attention: these techniques generally model the scene as a Markov random field (MRF) and include within a unique framework cues coming from local comparisons between the two images and from scene depth smoothness constraints. Another example of global method that works well is the one based on graph cuts.

Semi-global methods similarly to global methods adopt a global disparity model, but in order to reduce the complexity, minimization of the cost function is computed on a reduced model for each point of disparity image, differently than global approaches which estimate a whole disparity image at once. For example, the simplest semi-global methods such as Dynamic Programming or Scanline Optimization, work in a 1D domain and optimize each horizontal image row by itself. Semi Global Matching (SGM) algorithm is a more refined semi-global stereo algorithm and it will be presented next in Chapter 4.3

Disparity refinement or subpixel interpolation is typically obtained with an inexpensive technique like interpolating three matching costs with a parabola or splines centered on the minimum cost. Image filtering can also be used at some additional cost. Usually this is done, without enforcing any constraint about the underlining disparity map, by means of a median filter, morphological operators or bilateral filtering.

In the same article, Scharstein and Szelisky also propose a standard protocol for quantitative evaluation of a stereo algorithm. Their implementations and sample data, as well as an updated table with the performance of all the submitted algorithms, are available on the Web [23].

Although new algorithms improve more and more the solution of the correspondence problem, eventually the quality of stereo reconstruction intrinsically depends on the scene characteristics.

It is worth to mention that an active method can be used to reinforce the stereo matching computation, especially in uniform areas. With the aid of an external structured lighting device, two main approaches can be exploited. *Active stereo* involves still two cameras, and the external light is used to ease the correspondence selection. In *active triangulation* instead, just a camera and a calibrated projector

are used. Correspondence selection and triangulation procedure in this case have to be modified.

4.2 Practical issues

It can be argued that the detection of pairs of conjugate pixels is the most complex part of a depth map estimation. More generally, this is one of the major challenges in computer vision. Correspondence problem relies on the main assumption that left and right images are not too different, they have to exhibit a certain level of disparity while framing the same scene. Many problems afflict correspondence detection, mainly due to different perspective of the two cameras, and they get worse as the baseline increases. On the other hand a large baseline is needed to obtain a significative disparity. The major issues related to correspondence selection are now described.

Occlusions and discontinuities Due to discontinuities of the surfaces and particular displacement of the object in the scene, some points in one image may not exist in the other. For those points that do not have the relative conjugate, disparity has no reason and meaning to be defined. This is maybe the most known problem in stereo vision and can be observed by looking at the edge of an object first with one eye and then with the other: the background close to the edge is visible only with one of the two eyes. There exists a common procedure to detect occlusions, called *Left-Right consistency check*, but no solutions exist to retrieve the disparity of such areas.

Radiometric distortion and noise For materials not perfectly lambertian, the observed point can be different in the two images. Moreover due to the always present noise, color and intensity of the two acquired scenes can be different, increasing the complexity in the correspondence search.

Specular surfaces Similar to the previous issue, glossy materials may reflect the light directly into the camera. Due to different viewpoint of the two cameras, a region in one image may be visible and the correspondent in the other one may be overexposed. If the illumination of the scene does not come from a direct spot light, the likelihood of having such overexposed regions decreases.

Perspective foreshortening Because each stereo camera has a slightly different view, the image of the surface is more compressed and occupies a smaller area in one view. The more an object is horizontally slanted, the more pronounced

this effect is. Foreshortening causes problems especially to methods using fixed-size windows to aggregate costs, because they tacitly assume that objects occupy the same extents in both images.

Transparent objects Objects with a certain transparency present an intrinsic ambiguity. Background viewable through these objects actually would be occluded or even hided by it. This inevitably introduces an unwanted uncertainty that influences the results of both local and global methods.

Uniform regions Poor textured areas still continue to plague stereo matching systems. The ability to detect similar regions assumes that correlation or other methods are able to detect a peak of some functions. If a uniform region sufficiently large is considered, for example a white wall, neither local or global method can overcome this issue with sufficient certainty. Although this is a common problem in all stereo matching methods, techniques that propagate disparity cues are likely to assign a valid disparity also to these regions.

Repetitive pattern Correspondence of regions without texture is difficult to find, and so is the case of highly textured regions with periodic patterns. Without a global knowledge of the scene, it is impossible to distinguish between the correct correspondence or an erroneous translated version. A classic example is provided by framing a check board, in this case it is easily deductible that the cost function shape of the points inside the check board presents a certain number of peaks. Also in this case, the ambiguity can be reduced with the aid of global methods.

All these physical issues account for increasing the probability of false correspondence. Some of them can be handled by means of image processing or other techniques, but others, like occlusions, are physically impossible to manage. From this analysis, it can be argued that global methods solve many problems, improving the disparity estimation in regions where local methods fail, such as occlusions and uniform regions.

Correspondence problem is therefore afflicted by many practical issues, such a sought pair may not exist because of occlusions or perspective distortion and even if it exists it may not be straightforward finding it.

4.3 Depth estimation from stereo vision

Among the many algorithms available in the literature for disparity computation, the Semi-Global Matching (SGM) approach proposed by Hirschmuller [14] has been adopted in this thesis. 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. Authors propose to use 8 or 16 different independent paths. For each point, the disparity corresponding to the minimum aggregated cost is selected.

OpenCV [25], one of the most used library for real-time computer vision algorithms, provides an optimized implementation of a modified version of this algorithm. The only big difference is in the matching cost computation: Birchfield-Tomasi sub-pixel metric is used instead of the original mutual information cost function. Since they provide also the source code, this implementation has been adopted with some changes to extract intermediate results, essential information for confidence estimation. This section briefly reviews the main concepts of this modified algorithm.

The data fusion framework presented in Chapter 5 is independent of the choice of the stereo vision algorithm, therefore any choice is potentially suited to extract a disparity map and relative confidence estimations. A comparison on standard stereo matching algorithm [31] however shows that SGM is among the fastest methods and produces very good results, especially when efficiency is an issue.

In the following analysis, without loss of generality, it is assumed that the epipolar lines are parallel and horizontal, i.e. the stereo system is rectified. With this assumption the disparity map is therefore a scalar field: every pixel represents the horizontal shift between conjugates points.

The algorithm is described going through the four distinct processing step previously introduced.

Matching cost in the original paper is computed using a mutual information based approach for compensating radiometric differences of input images. In this implementation instead, the faster cost calculation provided by the sampling insensitive measure of Birchfield and Tomasi [2] is used. Another valid alternative is the census cost function, that gives the best overall results for different datasets and is rather robust under adverse lighting conditions. The cost is calculated as the absolute minimum difference of intensities at pixel $\mathbf{p}_L = [u_L, v_L]$ and the correspondent at pixel $\mathbf{p}_R = [u_R, v_R]$ in the range of half a pixel in each direction

along the epipolar line. Given I_L and I_R the intensity function of the two epipolar line in the two images, $\bar{I}_L$ and $\bar{I}_R$ the same function but up-sampled of a factor two, the dissimilarity $\mathcal{D}(\mathbf{p}_L)$ between $\mathbf{p}_L$ and $\mathbf{p}_R$ with respect to $\mathbf{p}_L$ is given by

$$\mathcal{D}(\mathbf{p}_L) = \min_u \left| I_L(u_L) - \bar{I}_R(u) \right| \ , \ u_R - \frac{1}{2} \leq u \leq x_R + \frac{1}{2} \tag{4.1}$$

where $u_R = u_L - d$. In the same manner the dissimilarity $\mathcal{D}(\mathbf{p}_R)$ with respect to $\mathbf{p}_R$ is

$$\mathcal{D}(\mathbf{p}_R) = \min_u \left| I_L(u) - \bar{I}_R(u_R) \right| \ , \ u_L - \frac{1}{2} \leq u \leq x_L + \frac{1}{2} \tag{4.2}$$

The cost $C(\mathbf{p}_L, d)$ of the disparity hypothesis d is defined as the minimum between $\mathcal{D}(\mathbf{p}_L)$ and $\mathcal{D}(\mathbf{p}_R)$.

Cost aggregation is the real strength of this approach. Pixelwise cost is generally prone to wrong matches, therefore an additional constraint is added to support smoothness and penalize changes of neighboring disparities. By assuming that the observed surfaces are quite smooth, disparity shifts can be penalized by setting an additional cost of assigning a depth to a pixel if it does not agree with its neighbors. This means that when the algorithm tries to estimate a point depth having several possible matches, it will probably choose the match which agrees more with the depth estimates of the neighboring pixels. The resulting energy function that depends on the disparity image D is defined as

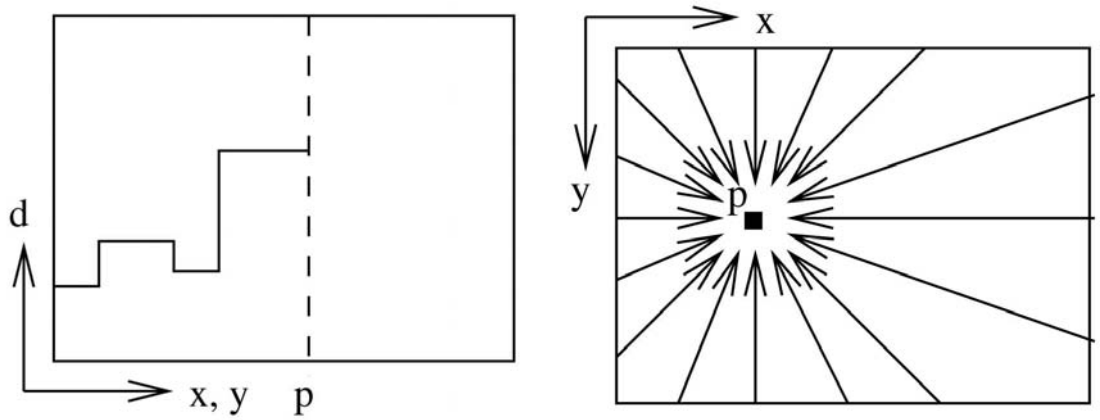
$$\begin{aligned} E(D) = \sum_{\mathbf{p}_L} \Big(C(\mathbf{p}_L, D_L) + & \sum_{q \in \mathcal{N}_{\mathbf{p}_L}} P_1 \ T[|D_{\mathbf{p}_L} - D_q| = 1] \\ & + \sum_{q \in \mathcal{N}_{\mathbf{p}_L}} P_2 \ T[|D_{\mathbf{p}_L} - D_q| > 1] \Big) \end{aligned} \tag{4.3}$$

where the first term accounts for pixel matching costs for the disparities of D, the second term adds a small penalty P_1 for all pixels in the neighborhood $\mathcal{N}_{\mathbf{p}_L}$ of $\mathbf{p}_L$ for which the disparity changes a little bit[1], and the last term adds a larger penalty P_2 ($P_2 \geq P_1$) for preserving discontinuities.

Unfortunately, a 2D global optimization of this energy function is $\mathcal{NP}$-complete. In contrast, 1D optimization can be performed efficiently in polynomial time. The idea of the authors involves searching in multiple directions to enforce a

[1] $T[A] = 1$ if the event A is true, 0 otherwise

global smoothness constraint on the solution. If this additional constraint was not considered, the disparity for each pixel would be computed without considering the estimated disparity of its neighbors, resulting in a noisy map with high probability of having many false positives. Searching in more directions increases the number of considered neighbors in the cost calculation and this will generally increase the likelihood of finding the correct disparity. Figure 4.1b shows an example of cost aggregation along 16 directions.



(a) Minimum cost path for a pixel $\mathbf{p}$ **(b)** 16 paths from all directions $\mathbf{r}$

Figure 4.1: Aggregation of costs in disparity space

The real implementation of these ideas requires to modify Equation (4.3) by defining the cost $L_{\mathbf{r}}(\mathbf{p}_L, d)$ along a path traversed in the direction $\mathbf{r}$ of the pixel $\mathbf{p}_L$ at disparity d

$$\begin{aligned} L_{\mathbf{r}}(\mathbf{p}_L, d) = C(\mathbf{p}_L, d) + \min \Big(& L_{\mathbf{r}}(\mathbf{p}_L - \mathbf{r}, d), \\ & L_{\mathbf{r}}(\mathbf{p}_L - \mathbf{r}, d-1) \ + \ P_1, \\ & L_{\mathbf{r}}(\mathbf{p}_L - \mathbf{r}, d+1) \ + \ P_1, \\ & \min_i L_{\mathbf{r}}(\mathbf{p}_L - \mathbf{r}, i) \ + \ P_2 \Big) \end{aligned} \tag{4.4}$$

In Figure 4.1a an example of minimum cost path $L_{\mathbf{r}}(\mathbf{p}_L, d)$ is depicted.

The number of paths must be at least 8 but with 16 a better coverage of the 2D image is provided and better performance is guaranteed. The final cost $C(\mathbf{p}_L, d)$ is defined as the summation along all the paths $\mathbf{r}$

$$C(\mathbf{p}_L, d) = \sum_{\mathbf{r}} L_{\mathbf{r}}(\mathbf{p}_L, d) \tag{4.5}$$

The *disparity image* $D_{\mathbf{p}_L}$ can be computed as the argument that minimizes (4.5), that is

$$D_{\mathbf{p}_L} = \arg\min_d C(\mathbf{p}_L, d) \tag{4.6}$$

However, a better performance can be achieved if the aggregate cost computation is performed also in the right image. A subsequent left-right consistency check on $D_{\mathbf{p}_L}$ and $D_{\mathbf{p}_R}$ enforces the uniqueness constraint, providing a better disparity estimation.

The resulting image may still contain some errors that can be removed with the *disparity refinement* step. Some of these improvements concern peaks removal and discontinuity preservation through a tailored interpolation technique.

The SGM approach works well in almost all scenarios, with good results also near depth discontinuities. However, due to its multiple 1D disparity optimization strategy, it produces less accurate results than more complex 2D global optimization approaches. This method is very fast and potentially capable to deal with poorly textured regions, thanks to the propagation of disparity hypothesis along multiple paths.

4.4 Confidence estimation of stereo disparity

The usage of confidence maps in applications based on stereo vision is mostly limited to remove those points with a low confidence, hence obtaining a sparse disparity map. This is straightforward but it does not take full advantage of the available information. Recently, stereo confidence computation has attracted rising attention and other useful applications have been devised thanks to this confidence knowledge. In [17] Hu and Mordohai present an extensive evaluation of confidence measures for stereo matching with the goal of detecting occluded points and of generating low-error depth maps by selecting among multiple hypothesis for each pixel. To this end, the disparity values are stored according to their confidence values, then, those depth measurements with the lowest confidence are dropped and a new error metric is calculated for the remaining pixels. The authors of [12] applied some confidence metrics to SGM. They put their efforts to reduce the number of not detected bad pixels and the number of discarded good pixels. The most recent contribution comes from a Daimler AG research [27]: the authors of this paper contributed to the first-time fully probabilistic usage of stereo confidences

along with the disparity map. They proved that instead of simply thresholding the disparity map, using confidences in a Bayesian manner yields a substantial improvement.

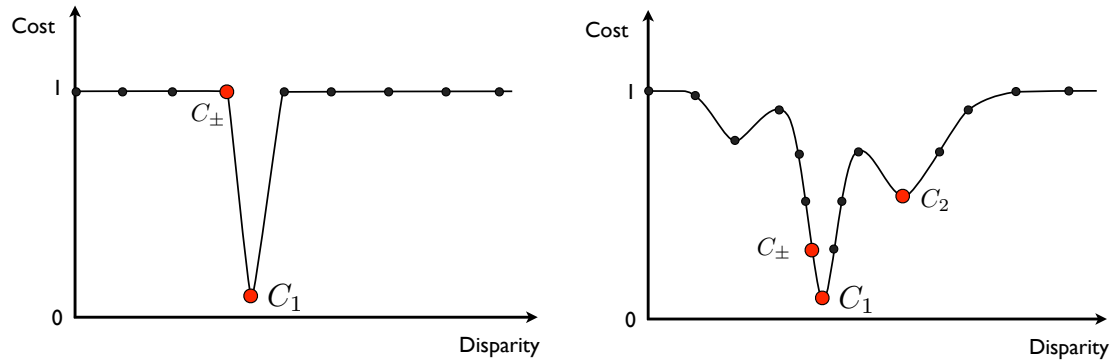
Among the plethora of confidence metrics defined over time, in the following only the most prominent measures in detecting wrong matches will be presented. An optimal confidence measure that aims at including all the properties of these maps is finally presented.

4.4.1 Cost curve analysis

In this thesis, the analysis is focused on individual pixels by examining their cost curves. The cost value assigned to a disparity hypothesis d for a pixel (u, v) is the one defined in Equation (4.5), and for this analysis will be denoted as $C(d)$, without the explicit pixel coordinates label as they are unambiguous. Moreover, the cost range has been normalized to the unit interval, i.e.

$$0 \leq C(d) \leq 1 \tag{4.7}$$

The ideal cost curve for a pixel, as a function of disparity, in Figure 4.2a is shown. The ideal cost is 0 for the correct disparity and 1 for all the others. It is reasonable to believe that if for a pixel the cost curve exhibits a behavior like the one depicted in Figure 4.2b, the disparity estimation will be more ambiguous. This is due to the multiple presence of local minima or multiple adjacent disparities with similar costs, making exact localization of the global minimum hard and often uncertain.



(a) Ideal cost **(b)** Ambiguous cost

Figure 4.2: Examples of cost curves

Figure 4.2b also shows the terminology used to denote some point of interest. The minimum cost for a pixel is denoted by C_1 and the corresponding disparity

value by d_1, i.e.

$$C_1 = C(d_1) = \min C(d) \tag{4.8}$$

The second smallest value of the cost that occurs at disparity d_2 is C_2. Very similar and adjacent to d_1 costs are excluded not to penalize disparity results around half integer values. For example, a second small cost can be considered valid if the distance of d_2 to d_1 is greater than 1. For some metric it is useful to define also $C_\pm$, associated to $d_\pm$, as the maximum between the two costs adjacent to the optimal disparity, i.e.

$$C_\pm = C(d_\pm) = \max(C_-, C_+) \tag{4.9}$$

where C_- and C_+ are the two costs adjacent to the optimal disparity cost C_1.

4.4.2 Peak Ratio Naive confidence

According to [17], Peak-Ratio Naive (PKRN) belongs to the category of confidences that consider local minima of the cost curve. The presence of other strong candidates is an indication of uncertainty. Different implementations of this metric can be considered, for example the simple Peak Ratio measure can be the ratio between the second smallest local minimum and the minimum cost C_1. A naive version of such ratio does not require the numerator to be a local minimum, indeed the second smallest cost C_2 can be used. This second definition assigns low confidence to matches with flat minima or strong competitors. The confidence measure actually implemented is

$$C_{PKRN} = \frac{C_2 + \varepsilon}{C_1 + \varepsilon} - 1 \tag{4.10}$$

where ε is a positive constant. This definition is slightly different from the original but offers some advantages. The minimum cost C_1 in rare cases may be 0, leading to singularity if ε is not considered, moreover this implementation is more robust to noise, especially at low cost levels. A typical value of ε is around 0.5, such values entail a limited dynamic range with a distribution rather uniform. The resulting confidence map has been normalized to the unit interval.

4.4.3 Maximum Likelihood Metric confidence

Following the categorization of [17], Maximum Likelihood Metric (MLM) belongs to the group of confidences that convert cost curve to a probability mass function over disparity. By assuming that the cost follows a normal distribution and that the disparity prior is uniform, after normalization, C_{MLM} is defined as

$$C_{MLM} = \frac{e^{-C_1/2\sigma^2}}{\sum_d e^{-C(d)/2\sigma^2}} \tag{4.11}$$

where σ represents the disparity uncertainty. Usually this value is chosen relatively high, e.g. $\sigma = 0.03$, to obtain a more uniform distribution. MLM has been classified as the second best method near discontinuities in detecting correct matches. It generates confidence maps with the sharpest boundaries.

Some variant of this method have been proposed, for example, a Gaussian distribution centered at the minimum cost value can be used. The resulting confidence map may not span the entire unit interval, therefore a normalization has been introduced.

4.4.4 Local Curve confidence

This confidence metric comes from [38] and exploits the Local Curve (LC) information of the equiangular fit. The shape of the cost curve around the minimum is an indicator of the quality of the match: a sharp valley indicates a good match, while flatness is an indication of uncertainty. This method is very similar to the curvature fit of parabola interpolation schemes. LC is computed as

$$C_{LC} = \frac{C_{\pm} - C_1}{\gamma} \tag{4.12}$$

where γ is a positive constant introduced to normalize the distribution. This variable can be avoided if a subsequent normalization is computed.

Curvature confidence has not been classified as one of the best methods for detection of correct matches according to [17]. It tends to rank some errors very highly because it assigns high confidence to pixels near discontinuities due to the related large discontinuities in the cost curve. However it comes with no additional computation as it is an intermediate result of the sub-pixel interpolation step.

4.4.5 Overall confidence

Most of the works on confidence estimation that have been analyzed usually provide a global likelihood measure by defining it as the product of some other metrics. This is the easiest way to combine different metrics, but it implicitly assumes independence among confidences. Confidence measures however are not pairwise independent: the correlation among them is quite strong as their definition is derived from the same cost function C.

After a careful analysis of the previous confidence metrics for different scenarios it was found that:

- When the minimum cost C_1 is above a certain threshold, for example in the top 25%

$$0.75 \leq C_1 \leq 1 \tag{4.13}$$

  then the associated disparity is with very high probability wrong, therefore a confidence of 0 is assigned to those pixels. Confidence estimation computed locally is not ideal for detecting global issues such as occlusions, since they are result of a long range interaction between surfaces. However it is customary to consider that a high matching cost is an indicator of occlusions or in general of not so reliable measures.

- Another clue of wrong disparity occurs when the width of the peak, or even multiple peaks, is larger than a certain threshold, that can be for example about a fifth of the entire disparity interval. Also in this case a confidence of 0 is assigned.

If the cost function behavior is not included into neither the first nor the second case, than to the estimated disparity can be assigned a confidence greater than 0:

- When the minimum cost is below a certain small value, for example

$$C_1 \leq 0.1 \tag{4.14}$$

  then the cost function is basically equal to the ideal shape of Figure 4.2a, therefore a confidence of 1 is assigned to those pixels.

- Otherwise, the last case is verified when the cost function has a minimum cost that is above that small threshold and the peak is relatively sharp. In this case a valid confidence value can be the difference between C_2 and C_1.

All these results are summarized in Figure 4.3.

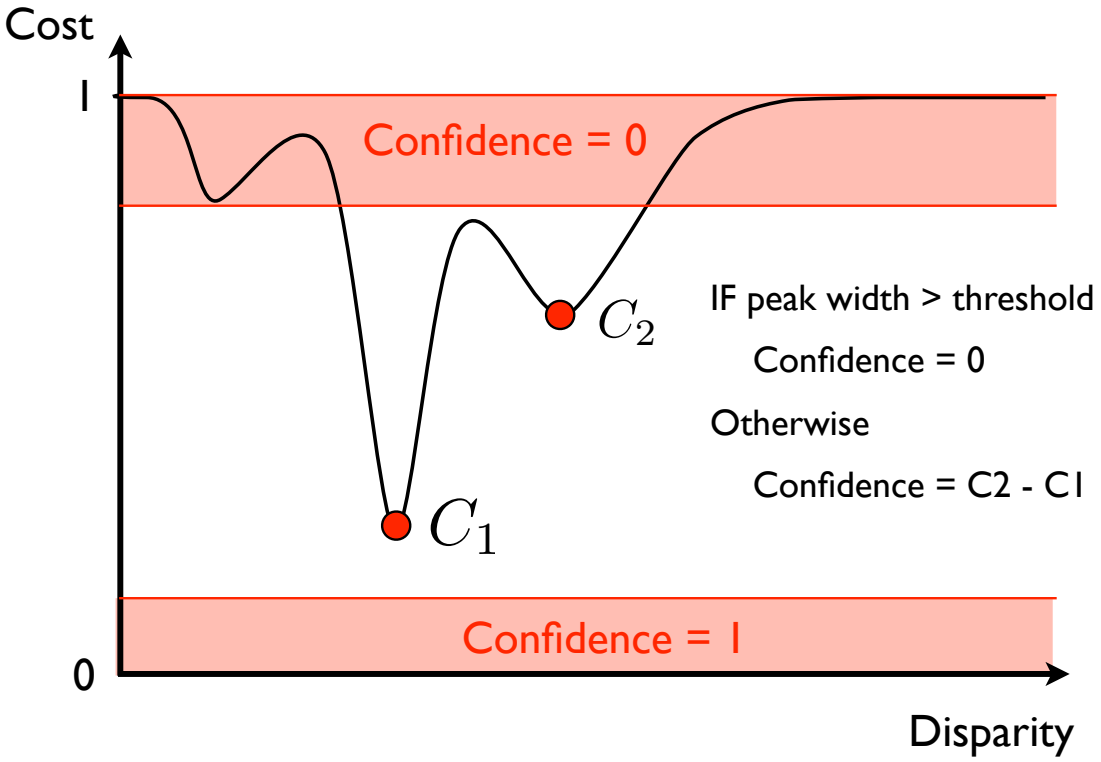


Figure 4.3: Characterization of cost function

This new confidence estimate has been inspired by the three measures previously introduced, Local Curve is another way to verify the peak width, Maximum Likelihood Metric and Peak Ratio Naive instead is replaced with the simplest difference, also known as Maximum Margin (MMN) metric. Implicitly also Matching Score Measure (MSM) is considered, and it associates a confidence value proportional to the minimum cost C_1, where low cost means high confidence, in this case equal to 1.

All these confidence estimations have been computed locally. Global methods in stereo vision generally are able to solve more practical issues. The same concept could be applied to confidence estimation, a global approach maybe could lead to better performance.

Chapter 5

Disparity map fusion

In the previous two chapters two different disparity maps have been computed, one from a ToF camera and one from a stereo system. Due to their complementary characteristics, it is reasonable to combine them in order to obtain a better disparity map. Theoretically, if someone would be able to label each pixel according to the disparity correctness, it would be sufficient to select, for each pixel, the correct hypothesis among the two provided. Unfortunately no information on disparity correctness is available, therefore the best way to discriminate among different hypothesis is to associate some kind of confidence information. In this chapter, a method to fuse the two disparity maps exploiting associated reliability information is described. The algorithm is an extension of the Local Consistency technique for cost aggregation, therefore after a discussion on the original implementation, the modified version will be presented.

5.1 Local Consistency technique

Local Consistency [19] is an approach devised to deal with classical problems of cost aggregation. The mutual relationships among neighboring points is exploited to derive a point based function that locally captures the global geometric and photometric structure of the scene. The goal of this algorithm is to improve the quality of a given disparity map by forcing the smoothness of the acquired scene, with the aid of additional color and spatial constraints to confine the smoothness hypothesis.

Considering Figure 5.1, the window around the red pixel represents the "global" vision of that pixel. With the green square a generic pixel inside the scope window is represented. The main idea of this algorithm is for each red pixel to propagate a disparity plausibility to all neighboring green pixels inside the window. The

plausibility of the disparity assumption for each element of the considered support can be modeled by these two events:

$\mathbf{E_1}$: this event encodes the belief that green pixels belong to the frontal support centered in the red pixel. Plausibility of this event is related to the color proximity between red and green pixels. Prior assumption that points closer to the central point are more relevant has been considered.

$\mathbf{E_2}(d)$: this event encodes the belief that green points in master and slave are homologous with disparity d and is related to the color proximity between them.

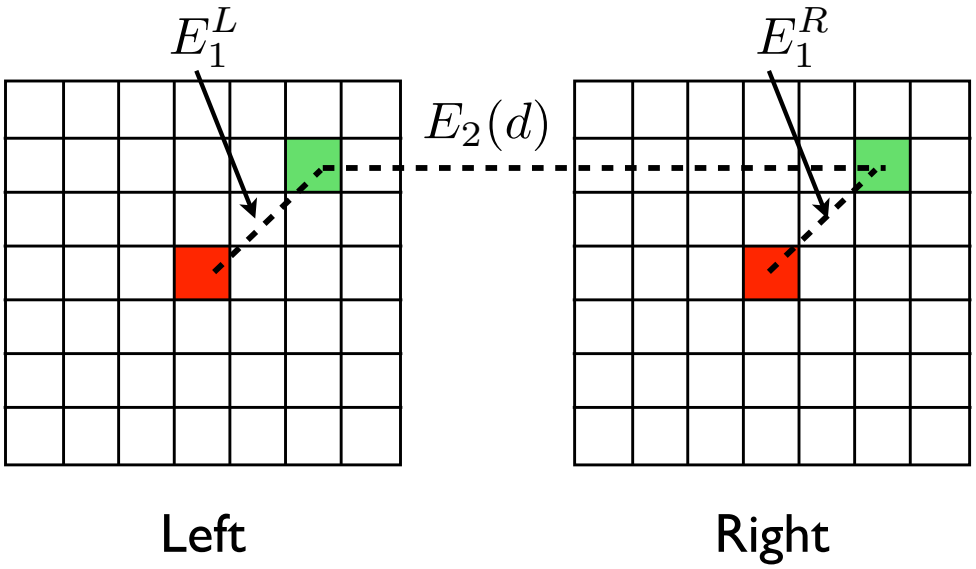


Figure 5.1: Events defining plausibility

Let Δ_C be a function that encodes color proximity in a certain color space, and Δ_D the Euclidean distance between red and green pixels. Plausibility is defined as the joint probability of the events depicted, given the spatial and color proximity

$$\mathrm{P}\left(E_1^L, E_1^R, E_2(d)^{LR} \mid \Delta_C^L, \Delta_C^R, \Delta_C^{LR}\right) \tag{5.1}$$

and thanks to Bayes' rule and events independence

$$\begin{aligned} \mathrm{P}\Big(E_1^L, E_1^R, E_2(d)^{LR} \mid \Delta_C^L, \Delta_C^R, \Delta_C^{LR}\Big) \propto \\ \mathrm{P}\Big(E_1^L\Big) \cdot \mathrm{P}\Big(\Delta_C^L, \mid E_1^L\Big) \\ \mathrm{P}\Big(E_1^R\Big) \cdot \mathrm{P}\Big(\Delta_C^R \mid E_1^R\Big) \\ \mathrm{P}\Big(E_2(d)^{LR}\Big) \cdot \mathrm{P}\Big(\Delta_C^{LR} \mid E_2(d)^{LR}\Big) \end{aligned} \tag{5.2}$$

where the first term of each row represent the prior, and the second represent the likelihood.

Prior for events E_1 has been set according the following spatial proximity constraint

$$\mathrm{P}\Big(E_1\Big) = e^{-\Delta_E/\gamma_E} \tag{5.3}$$

while no prior knowledge has been assumed for E_2. Assuming Δ_C Gaussian distributed, the overall plausibility will be

$$\begin{aligned} \mathrm{P}\Big(E_1^L, E_1^R, E_2(d)^{LR} \mid \Delta_C^L, \Delta_C^R, \Delta_C^{LR}\Big) \propto \\ e^{-\Delta_E^L/\gamma_E} \cdot e^{-\Delta_C^L/\gamma_C} \cdot e^{-\Delta_E^R/\gamma_E} \cdot e^{-\Delta_C^R/\gamma_C} \cdot e^{-\Delta_C^{LR}/\gamma_T} \end{aligned} \tag{5.4}$$

where γ_i are parameters to control the behavior of such distributions. For the sake of clarity, Equation (5.4) for pixel $\mathbf{p}$ will be denoted as $\mathcal{P}_{\mathbf{p}}(d)$.

For each pixel in the image, each pixel in the active window will receive a plausibility of a certain disparity d. With the reference of Figure 5.2, the red pixel will receive a plausibility from a certain number of pixels within the active window. In the considered example, green pixels have all the same disparity d and therefore will propagate a non null plausibility for d.

The overall plausibility for the red pixel at disparity d will be therefore

$$\Omega_{\mathbf{p}}(d) = \sum_{\mathbf{q} \in \mathcal{A}} \mathcal{P}_{\mathbf{q}}(d) \tag{5.5}$$

where $\mathcal{A}$ is the active window. This aggregation is computed both on master and slave images and then the results are normalized over the plausibility at all disparity levels.

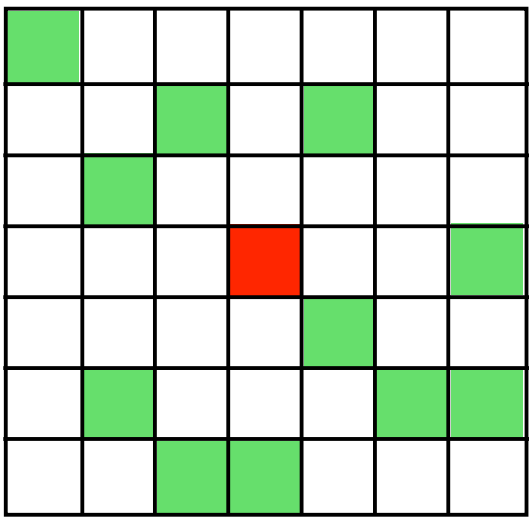


Figure 5.2: Plausibility accumulation

After these calculations, in order to obtain a robust disparity estimation, cross-validation of the accumulated plausibility has been computed

$$\Omega_{\mathbf{p}}(d)^{LR} = \Omega_{\mathbf{p}}(d)^{L} \cdot \Omega_{\mathbf{p}}(-d)^{R} \tag{5.6}$$

and then the disparity map $D(\mathbf{p})$ will be

$$D(\mathbf{p}) = \arg\min_{d} \Omega_{\mathbf{p}}(d)^{LR} \tag{5.7}$$

Hypothesis propagation allows to overcome many of the problems typical of local approaches, however the presence of wrong disparity hypothesis, e.g. due to occlusions, may perturb the aggregated plausibility. A left-right consistency check before running the algorithm may be useful to limit such undesired effects. The effectiveness of this algorithm is also visible if a sparse disparity map is used as input. Disparity propagation acts like an interpolating function, assigning (hopefully) valid disparity also to regions without original values. It is worth to notice that the plausibility function defined on color and range information ensures robustness to this approach at the cost of having multiple parameters that require an empirical estimation.

5.2 Modified Local Consistency for depth fusion

The extension to the case with two input disparity maps is pretty straightforward. The overall algorithm is exactly the same of the original implementation, except for the accumulated plausibility of Equation (5.5). The new formula is modified as follow

$$\Omega'_{\mathbf{p}}(d) = \sum_{\mathbf{q}\in\mathcal{A}} \left(\mathcal{P}'_{\mathbf{q},T}(d) \ + \ \mathcal{P}'_{\mathbf{q},S}(d) \right) \tag{5.8}$$

where $\mathcal{P}'_{\mathbf{q},T}(d)$ and $\mathcal{P}'_{\mathbf{q},S}(d)$ are the new plausibility of ToF and stereo cameras. In this new scenario, for each point of the input image there can be 0, 1 or 2 disparity hypothesis. If both sensors do not have a potential valid range measurement, no disparity cue is propagated; when only one of the two sensors has a potential valid range measurement, that value is propagated exactly as in the original algorithm. In the optimal case when both the disparity fields have a potential valid disparity value for a pixel, two cues will be propagated within the active support.

If Equation (5.4) is used as plausibility, all the cues would be propagated with the same weight. However an erroneous disparity hypothesis from a sensor could negatively impact the overall result. The introduction of a suited weight allows instead to discriminate between the (hopefully) two hypothesis. A reasonable choice is to define the new weighted plausibilities as

$$\begin{aligned} \mathcal{P}'_{\mathbf{q},T}(d) &= C_T(\mathbf{p}) \ \cdot \ \mathcal{P}_{\mathbf{q},T}(d) \\ \mathcal{P}'_{\mathbf{q},S}(d) &= C_S(\mathbf{p}) \ \cdot \ \mathcal{P}_{\mathbf{q},S}(d) \end{aligned} \tag{5.9}$$

where $C_T(\mathbf{p})$ and $C_S(\mathbf{p})$ are the confidences of ToF and stereo disparity measures respectively, and $\mathcal{P}_{\mathbf{q},T}(d)$ and $\mathcal{P}_{\mathbf{q},S}(d)$ as defined in (5.4). The adoption of this model for the new plausibility is supported by the nature of the confidence maps, indeed, such values can be interpreted as the probability that the corresponding disparity measure is correct. A confidence of 0 means that the disparity value is not reliable, then it is justified to not propagate such hypothesis. The opposite case is when the confidence is 1, that means for the associated disparity high likelihood of being correct. All the intermediate values will contribute as weighting factors. This definition is also coherent when a disparity value is not available, for example due to occlusions: the associated confidence is 0, therefore no values will be propagated.

It is worth to mention that another natural definition of the weighting factor

is an exponential function of the confidence, this can be thought as an extension of the plausibility function for uniformity with the likelihood terms. However the Gaussian assumption of each term in (5.4) is not valid in general for the confidence measure, since it has been defined as a point-wise likelihood indication, therefore this possibility has not been considered.

An interesting observation on the effectiveness of this framework is that equation (5.8) can be extended to deal with more than two input disparity maps simply adding other plausibility terms for the new disparity cues.

Chapter 6

Results

In order to evaluate the performance of the proposed fusion framework, a C++ program has been developed with the aid of the OpenCV library. This software allows to easily test the effects that different confidence metrics have on the fusion algorithm. Its modular structure allows also to extend the data fusion to more than two input maps by just implementing few methods of a common interface, or to replace the stereo matching algorithm to test the impact of different methods. All the parameters are read from an external file, allowing thus rapid experimentation.

For a robust analysis, it is important to have real-world data, therefore a dataset of 5 different static scenes has been acquired. The process of collecting data from stereo and ToF systems together requires long time and high accuracy, calibration in particular is the most expensive procedure. This is because depth estimation and data fusion results strongly depend on how the two systems are calibrated, therefore multiple attempt are needed to find the calibration parameters that minimize a certain error.

Benchmarks that compare stereo algorithms on a dense level are available for example at the Middlebury Stereo Vision Page [23] together with a complete set of images, calibration parameters and ground truth. However, a complete dataset with associated Time-of-Flight measures is still missing. One way to overcome this lack has been recently proposed in [20]: it consists on the synthetic generation of ToF data accounting for all scene-dependent effects.

Accuracy of the proposed framework has been evaluated by computing the mean squared error (MSE) of the resulting fused depth map with the ground truth previously computed.

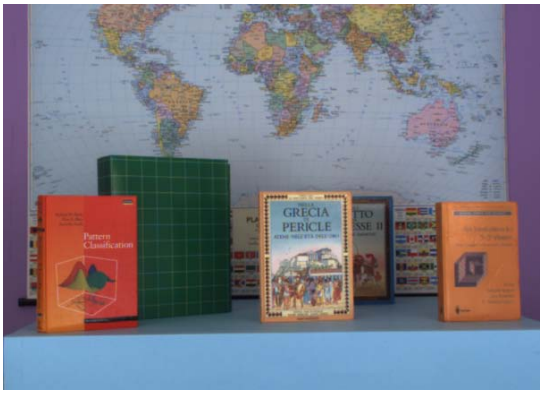
6.1 Dataset acquisition

Referring to Figure 2.3, the ToF sensor used for the dataset acquisition in this thesis is the MESA SwissRanger SR4000, with a 10 mm optics and horizontal field of view of 43°. It acquires a 16-bit depth image with values in $[0, 5\,\text{m}]$, a 16-bit signal amplitude image and a 16-bit confidence map. All these data are framed with a resolution of 176 x 144. The stereo vision system is made of two standard BASLER scA1000$^{\text{TM}}$RBG camera with 4.5 mm optics that acquire RGB images with resolution 1032 x 778. The baseline for the stereo pair is 170 mm. The overall system has been calibrated with the method proposed in [7], with a resulting spatial error of about 5 mm. The result of the fusion framework is a disparity map of the same resolution of the stereo system, i.e. 1032 x 778.

The five acquired datasets involve indoor scenario and external illumination sources chosen accurately to avoid interference with the ToF measurements. Due to the complementary characteristics of the two sensors, particular attention has been given in the scene arrangement. Dataset 2 presents piecewise smooth surfaces, ideal for the implicit assumption of stereo matching, but reflective materials and texture-less regions. Dataset 5 instead presents a more complex scene, with less reflective materials but with high textured areas. The other three datasets have intermediate characteristics, combining depth discontinuities, materials with different reflectivity and textured objects. The five acquired scenes from the left camera point of view, after undistortion and rectification, in Figure 6.1 are shown. The actual images acquired by the camera are bigger than the ones shown, however it is reasonable to evaluate and compare result in the region framed by both stereo and ToF cameras.

The disparity maps of Time-of-Flight and stereo vision system have been computed as described in Chapter 3 and 4. In Figure 6.2, the second column shows the disparity map obtained from the novel interpolation technique and the third column shows the disparity map for stereo vision provided by the modified Semi-Global matching algorithm. Disparity images have been represented with a classical Jet colormap to represent different values. The dark blue regions represent occlusions detected by the Left-Right consistency check, or region where a disparity value has not been assigned or is out of the predefined limits. It is interesting to notice that while those regions for stereo represent a relatively high percentage of the image, ToF disparity does not suffer from this problem, also due to the interpolation scheme.

Ground truth has been estimated with the spacetime stereo method described in [41]. By combining both spatial and temporal appearance variation, this approach reduces ambiguity and increases accuracy. A set of 600 images with 600 different



(a) Dataset 1

(b) Dataset 2

(c) Dataset 3

(d) Dataset 4

(e) Dataset 5

Figure 6.1: Acquired datasets

patterns have been acquired and combined to obtain a more accurate disparity map, sub-pixel refinement and Left-Right check also increase the accuracy. The precision of the depth maps obtained with such a system is of approximately $1-2$ mm. Since the ground truth has been obtained from a stereo vision procedure, not all the pixels will have a valid disparity due to occlusions. The last column of Figure 6.2 shows the 5 ground truth disparity maps.

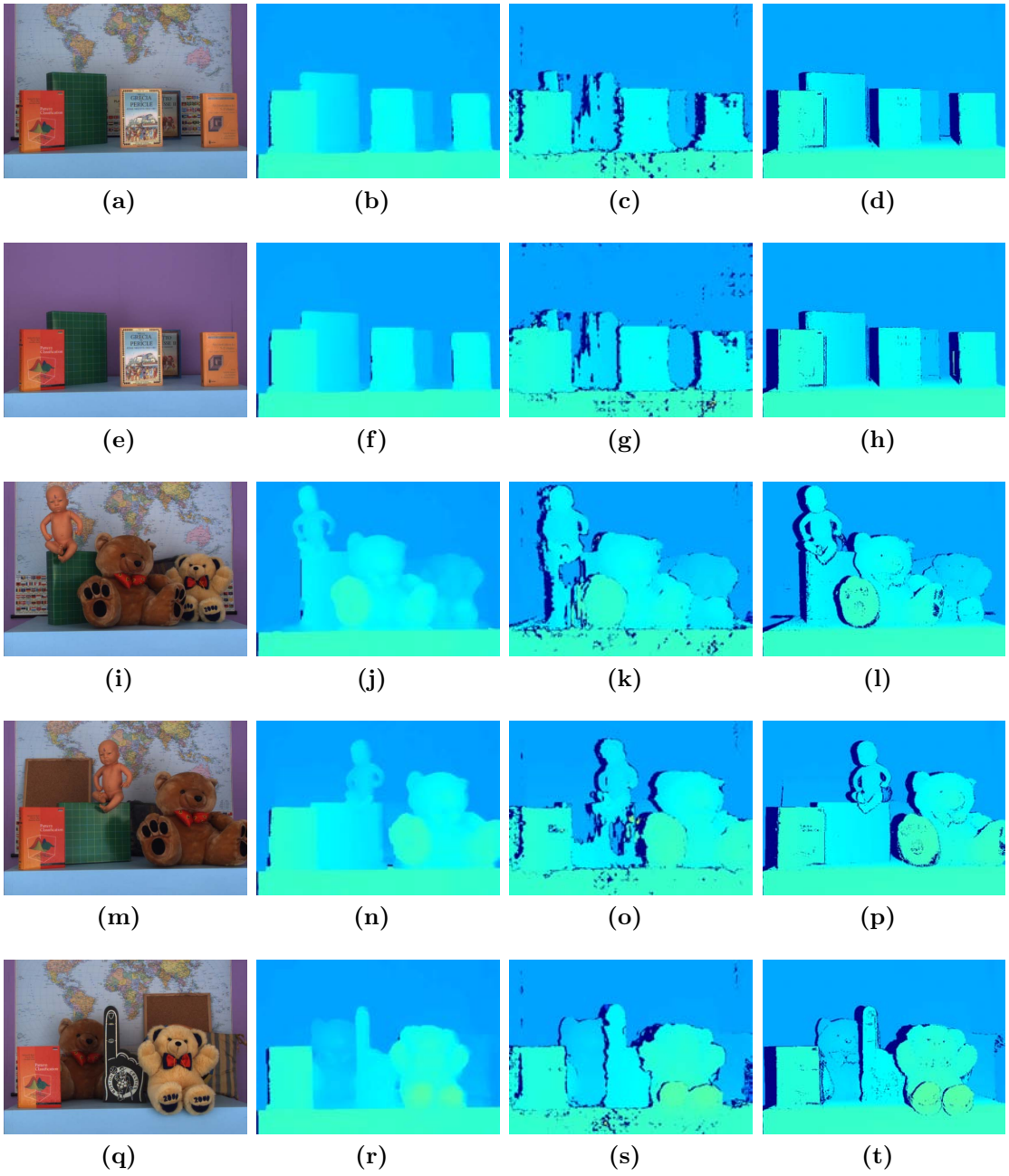


Figure 6.2: Disparity maps of Tof, stereo and ground truth

6.2 Confidence maps

In Chapter 3 and 4 some confidence maps have been presented. The purpose of a confidence measure is to assign a likelihood to each disparity measure. Therefore a good confidence map should be high correlated with the actual error, i.e. low confidence should be assigned to points with a disparity value different from the one of the ground truth and high confidence should be assigned to points with a correct disparity.

All the confidence maps associated to ToF disparity measures are depicted in Figure 6.3. The five columns show likelihood measures previously derived, from the first to the last column respectively: the confidence associated to the signal amplitude, the confidence map provided by ToF sensor, the combination of amplitude and intensity, the confidence from local variance and the product of the third and the fourth confidence measures.

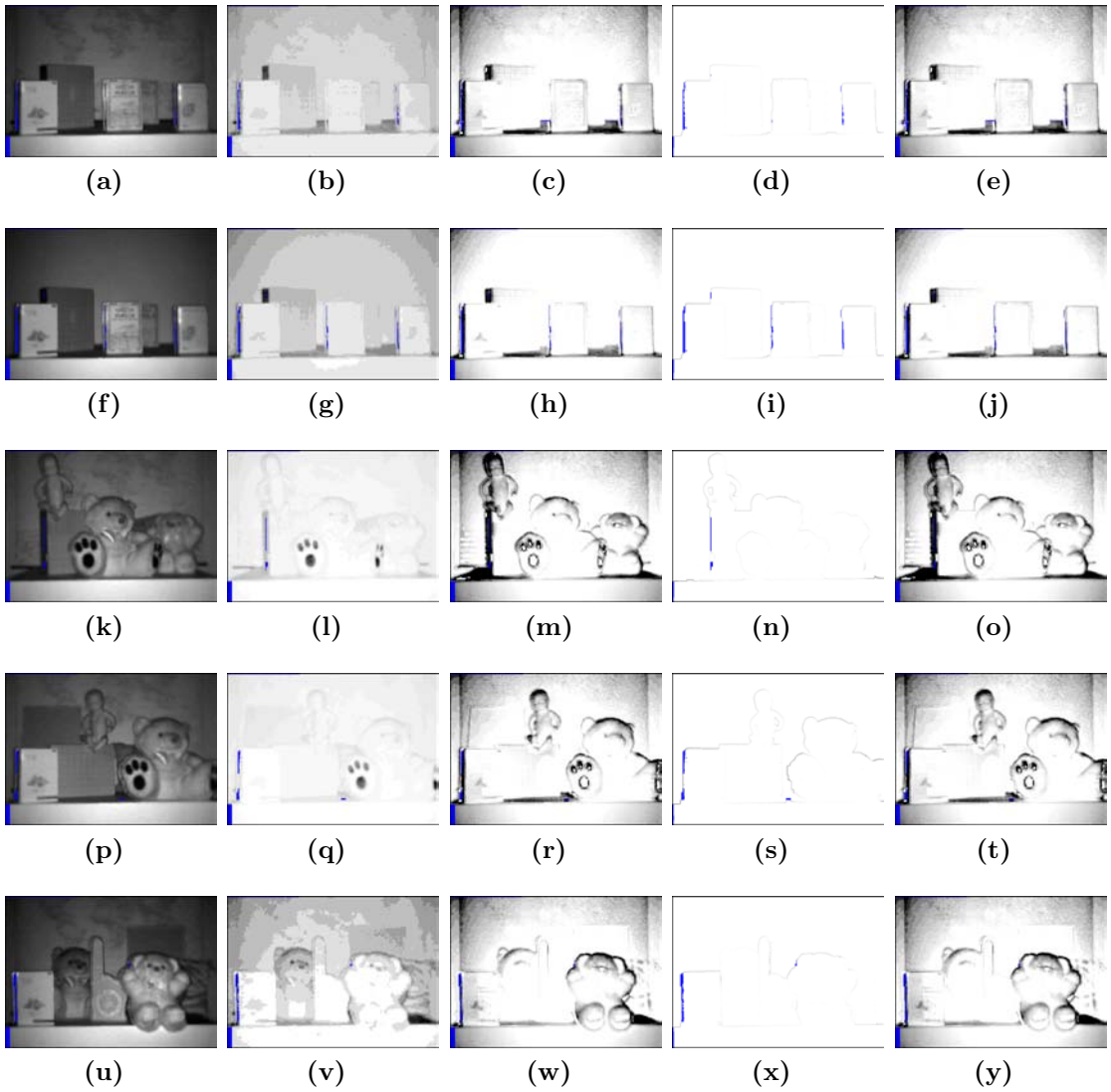


Figure 6.3: Confidence maps for ToF disparity

At a glance it can be noticed that, differently from the other methods, local variance (column 4) assign a confidence of almost 1 to all the points but the edges, exactly as expected. A common characteristic of all the other metrics is that they tend to assign lower confidence to the upper part of the table. This is as expected, as it is almost parallel to the emitted rays and the amplitude of the received signal is low. In the confidence provided by ToF camera (column 2), especially in the first

two datasets, the dependency of the likelihood from the distance to the center of the image is quite visible. This effect strongly influences also the confidence from amplitude and intensity (column 3) even if this metric has not been computed from confidence of ToF camera. In this third metric, regions close to the four corners of the image have always likelihood values close to 0.

These images confirm the observations that have led to the definition of the overall confidence. Amplitude of the received signal (column 1) is not sufficient to guarantee a robust confidence measure: some regions with this confidence value close to 0 become more reliable if also the intensity is considered. For example the dark circles in the foot of the teddy bear are ranked low confidence from the first measure (k), while in the third measure (m) they receive a higher confidence. A rapid comparison with Figure 6.6j, showing the mean square error with the ground truth, confirms that those regions should actually receive a high confidence. The multiplication by the confidence from local variance is only necessary to decrease the likelihood of edges.

For stereo disparity, four confidence metrics have been considered, and are depicted in Figure 6.4. From the first to the last column, the confidence metrics are respectively: Peak Ratio Naive, Maximum Likelihood Metric, Local Curve and the overall confidence.

From these images, it can be noticed that all the metrics assign generally high likelihood values but the Local Curve method (column 3). As previously discussed, Local Curve performs worse than expected given its popularity, and this is because if the cost peak is not really sharp, its curvature will be rather high. If for example Figure 6.4s is considered, it is easy to see that a lower confidence is associated to the book on the left, with respect to the other three metrics. A comparison with the mean square error image in Figure 6.6s however reveals that the estimated disparity is correct and thus a higher confidence should be assigned. In addition, this method experiences some problem also with repetitive patterns.

Maximum Likelihood Metric (column 2) assigns in all the datasets values very high at almost all the pixels, also where they should clearly be low. If for example the second dataset is considered, the disparity of the violet wall behind the table is for sure not precise given its color uniformity. However, from Figure 6.4f it is clear that the confidence associated to the wall region is not so different to the one associated to the other objects in the scene.

Peak Ratio Naive (column 1) and the proposed overall confidence (column 4) metrics perform similar, but the latter assigns a lower confidence to regions where clearly stereo information is not reliable, like for example regions with low texture.

A particular problem common to all these metrics based on cost curve can

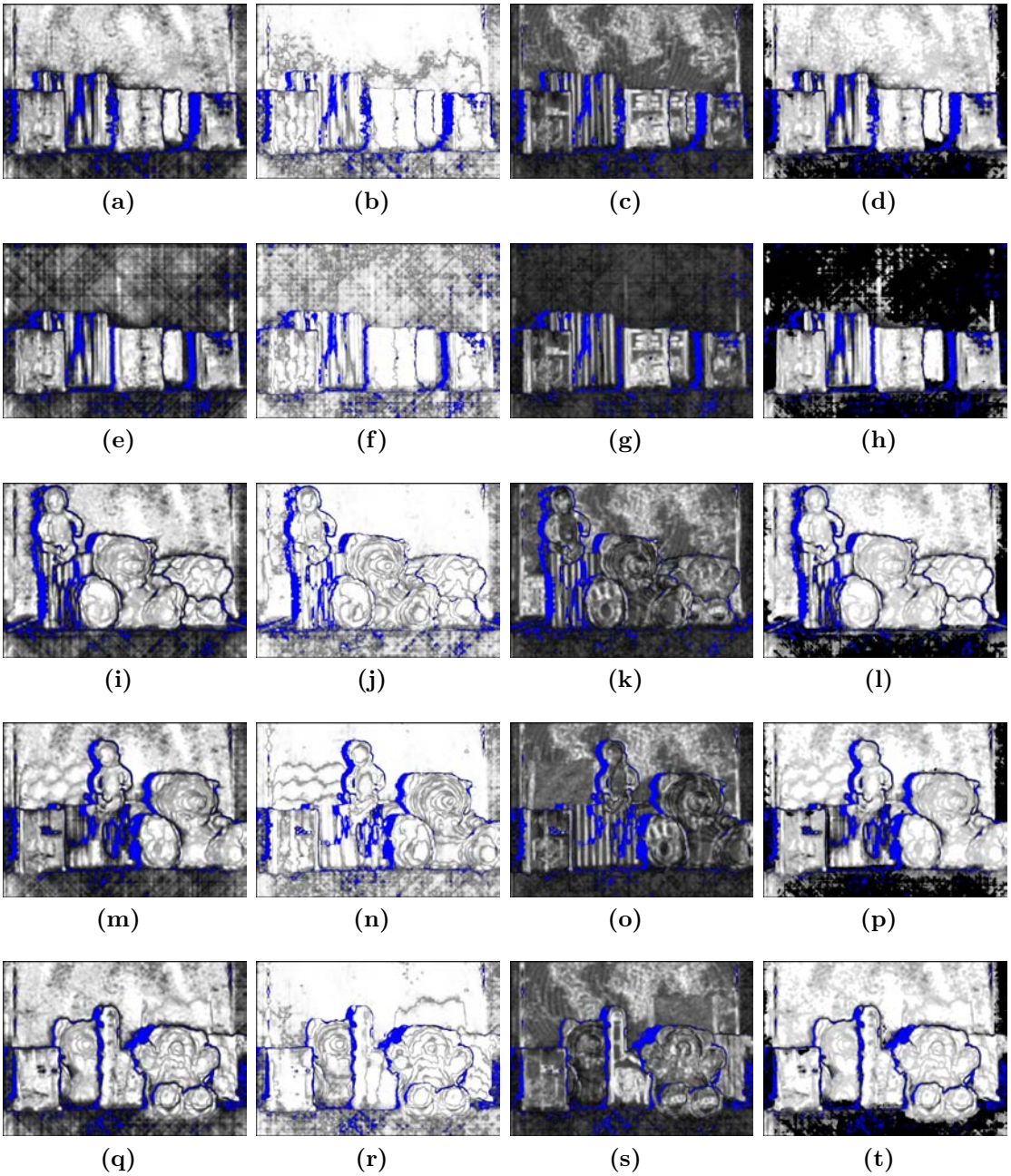


Figure 6.4: Confidence maps for stereo disparity

be noticed by looking at the estimated disparity map in Figure 6.2c. The upper part of the green book reveals that a big portion has completely wrong disparity values, therefore the associated confidence should be low. However all the four confidence metrics associate a relatively high value to this region and this is due to the particular shape of the cost function. A visual inspection of the cost curve of these points shows that its behavior is almost equal to the ideal shown in Figure 4.2a. For such cases, no confidence metrics based only on matching a cost function could reveal a wrong match.

6.3 Disparity fusion

The fusion framework takes as input the two disparity maps and the associated confidence measures, and gives as output a disparity map of the framed scene from the point of view of the left camera. If the proposed fusion approach was correct, the resulting disparity estimation should be better than the two obtained from ToF and stereo vision considered separately.

The metric used to compare the fusion results is the Mean Squared Error (MSE) between the estimated disparity and the ground truth. Points without a valid ground truth or without a valid estimated disparity are not considered in MSE computation. Different results in terms of MSE could be obtained if the depth maps were compared instead. All the combination of confidence metrics have been tested and the average MSEs of the five scenes have been compared.

The overall best performance was obtained with the two expected metrics, i.e. the so called overall confidence: for Time-of-Flight the combination of amplitude, intensity and local variance, and for stereo system the proposed metric. Figure 6.5 shows the optimal disparity maps (column 2) obtained with these confidence metrics and the relative ground truth (column 3). Numerical results of the MSE are instead listed in Table 6.1.

	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Average
ToF	3.271	3.255	4.128	3.373	4.130	3.631
Stereo	5.020	5.921	4.687	5.625	4.508	5.152
Fusion	**2.931**	**3.051**	**3.316**	**2.797**	**3.177**	**3.035**

Table 6.1: MSE with optimal confidence maps

From the numerical comparisons it can be noticed that, in general, the ToF camera provides better results, also thanks to the interpolation algorithm used. Moreover, errors due to low reflective materials or light reflection are limited. However, the lack of high textured regions cause problems to the stereo matching, explaining the high MSE values. Dataset 2, with planar object and reflective materials without high textured regions, is the ideal scene for the ToF and the worst ont for the stereo. Dataset 5, conversely, provides high textured and less reflective materials, resulting the worst scene for the ToF and the best one for the stereo. For all the five scenes the fusion MSE is lower than both ToF and stereo MSEs, therefore the weighted combination of the two hypothesis always overcomes the separated estimation.

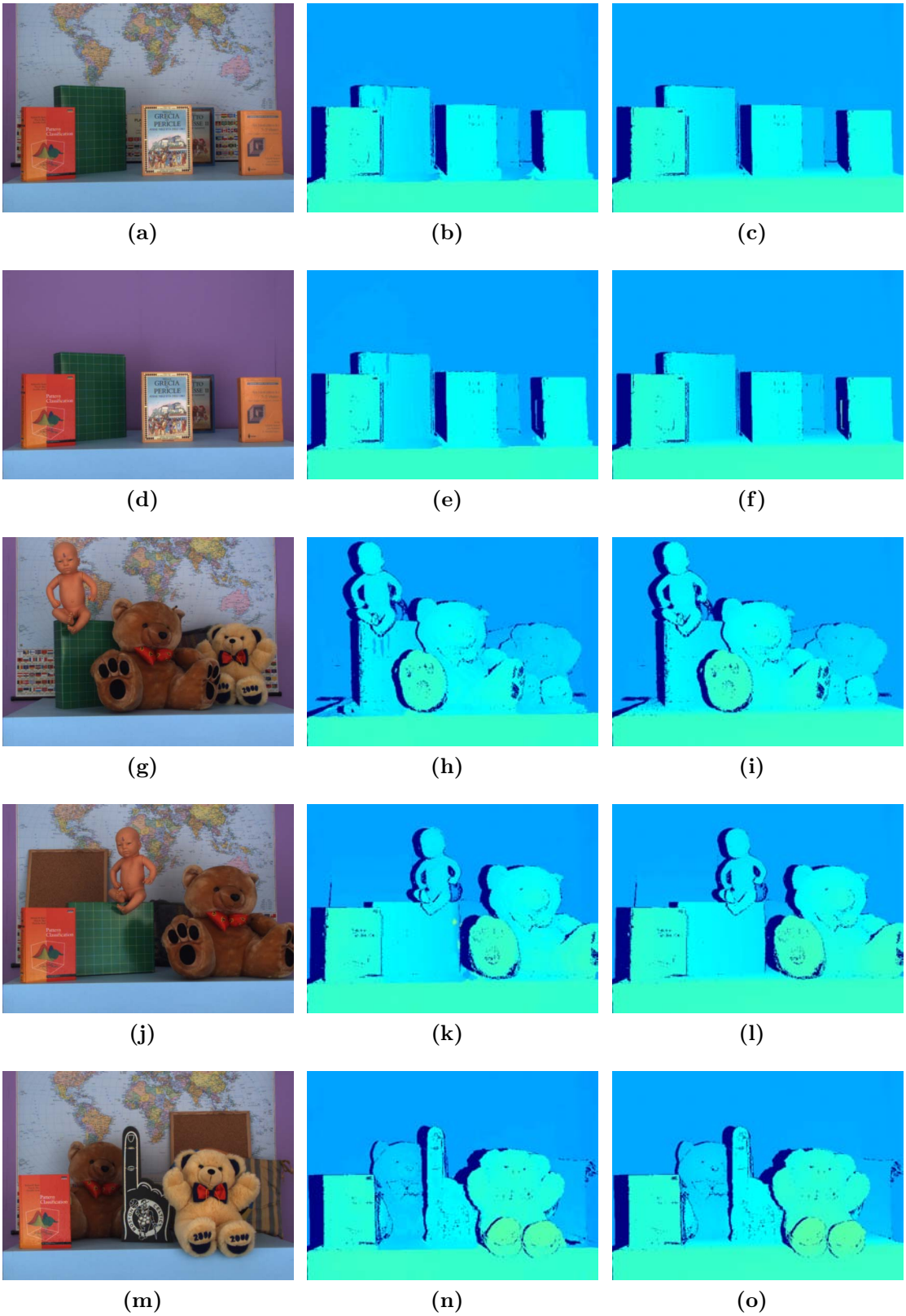


Figure 6.5: Disparity of the fusion algorithm with optimal confidence maps

Visual comparison of MSE allows to better understand results of the fusion algorithm. Figure 6.6 shows the five datasets and the associated MSE of the different disparity maps: the second column is relative to the ToF disparity, the third to the stereo vision and the last to the fused disparity map.

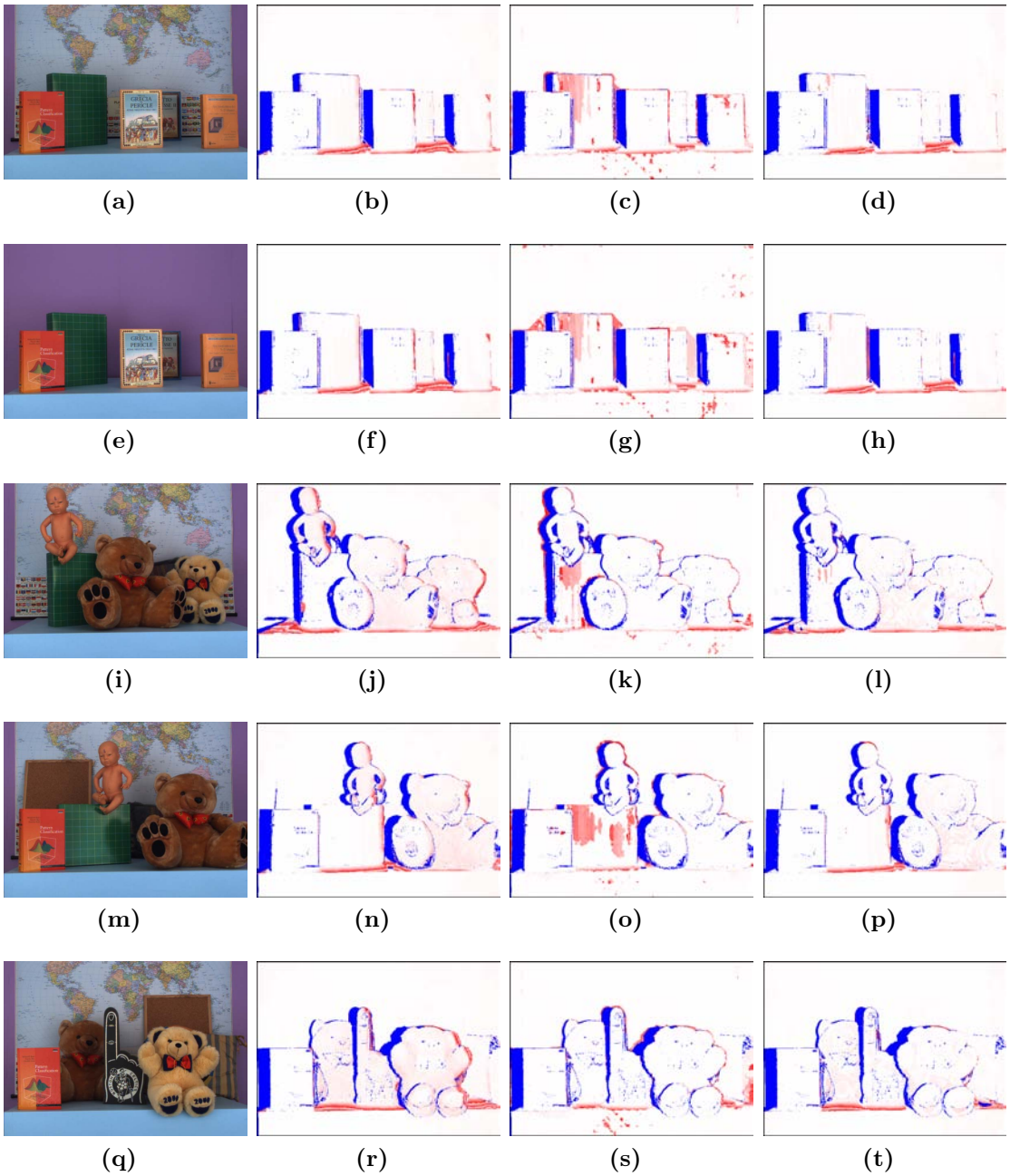


Figure 6.6: MSE images with optimal confidence maps

The results from the fusion algorithm (column 4) are generally good, the adopted confidence maps have led to select with high probability the correct disparity hypothesis from the two provided for each pixel. In addition, the Locally Consistent approach of the fusion algorithm allows to obtain a better estimation by propagating

disparity cues.

An erroneous disparity value in the upper region of the green book is almost always selected, even if the ToF camera provides valid disparity values for such area. This is due to the previously described issue with stereo matching cost in that region: the cost function presents a single sharp peak and the associated confidence is therefore high. However thanks to Local Consistency these effects have been mitigated.

Finally, it is worth to notice that residual errors still present in the fusion disparity are mainly due to the lack of a correct disparity in both the hypothesis.

Chapter 7

Conclusions

In this thesis, a framework for 3D data fusion with confidence information has been presented. The complementarity on the nature of data acquired by a Time-of-Flight camera and a stereo vision system suggests that a combination of these information might lead to performance improvement. Experimental results show that certain confidence measures, together with the particular fusion algorithm based on Local Consistency, actually provide a substantial enhancement of depth accuracy. For ToF measures, the best confidence metric exploits amplitude of the received signal, illumination intensity of the scene and local variance, while for stereo vision it has been found that the best confidence metrics do not require the knowledge of the overall cost function but just of the minimum and second smallest costs.

Only five different scenes have been considered to assess the quality of the proposed method, therefore the next step will be to evaluate the fusion algorithm in a bigger dataset, also introducing more variability in the scenes. A definition of a complete dataset with both stereo and ToF calibrated data together with the ground truth is for sure a fundamental requirement if results of data fusion have to be compared.

Another possible extension to this work is to explore the fusion of depth maps over dynamic scenes. ToF are sensitive to motion, therefore a confidence measure that takes into account temporal variation should be considered. Authors of [42] show how fusion techniques can benefit from the inclusion also of temporal domain, generating improved depth maps for dynamic scenes and therefore leading to significant improvement with ToF sensors.

Depth fusion from different sensors has recently attracted also commercial products like Microsoft Kinect, therefore fast developments must be expected in this field in the next years.

Bibliography

[1] *Basler*. URL: `http://www.baslerweb.com` (cit. on p. 27).

[2] S. Birchfield and C. Tomasi. "Depth discontinuities by pixel-to-pixel stereo". In: *Computer Vision, 1998. Sixth International Conference on*. 1998, pp. 1073–1080 (cit. on p. 32).

[3] G. Bradski and A. Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*. Cambridge, MA: O'Reilly, 2008 (cit. on p. 10).

[4] B. Buttgen and P. Seitz. "Robust Optical Time-of-Flight Range Imaging Based on Smart Pixel Structures". In: *Circuits and Systems I: Regular Papers, IEEE Transactions on* 55.6 (2008), pp. 1512–1525 (cit. on pp. 16, 22).

[5] B. Büttgen, 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: *In Proceedings of 1st Range Imaging Research Day* (2005), pp. 21–32 (cit. on pp. 16, 22).

[6] D. Comaniciu and P. Meer. "Mean shift: a robust approach toward feature space analysis". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 24.5 (2002), pp. 603–619 (cit. on p. 18).

[7] C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo. "A Probabilistic Approach to ToF and Stereo Data Fusion". In: *3DPVT*. Paris, France, May 2010 (cit. on pp. 10, 48).

[8] C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo. *Time-of-Flight Cameras and Microsoft Kinect(TM)*. Springer Publishing Company, Incorporated, 2012 (cit. on pp. 8, 10, 12).

[9] C. Dal Mutto, P. Zanuttigh, S. Mattoccia, and G. Cortelazzo. "Locally consistent tof and stereo data fusion". In: *Proceedings of the 12th international conference on Computer Vision - Volume Part I*. ECCV'12. Florence, Italy: Springer-Verlag, 2012, pp. 598–607 (cit. on pp. 3, 17, 19).

[10] V. Garro, C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo. "A Novel Interpolation Scheme for Range Data with Side Information". In: *CVMP*. 2009.

[11] S. Guomundsson, H. Aanaes, and R. Larsen. "Environmental Effects on Measurement Uncertainties of Time-of-Flight Cameras". In: *Signals, Circuits and Systems, 2007. ISSCS 2007. International Symposium on*. Vol. 1. 2007, pp. 1–4 (cit. on p. 14).

[12] R. Haeusler and R. Klette. "Optimality in combinations of confidence measures for stereo vision". In: *Proceedings of the 27th Conference on Image and Vision Computing New Zealand*. IVCNZ '12. New York, NY, USA: ACM, 2012, pp. 150–155 (cit. on p. 35).

[13] K. He, J. Sun, and X. Tang. "Guided Image Filtering". In: *Computer Vision – ECCV 2010*. Ed. by K. Daniilidis, P. Maragos, and N. Paragios. Springer Berlin Heidelberg, 2010, pp. 1–14 (cit. on p. 28).

[14] H. Hirschmuller. "Stereo Processing by Semiglobal Matching and Mutual Information". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 30.2 (2008), pp. 328–341 (cit. on p. 32).

[15] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz. "Fast Cost-Volume Filtering for Visual Correspondence and Beyond". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 35.2 (2013), pp. 504–511 (cit. on p. 28).

[16] R. Hrtley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004 (cit. on p. 9).

[17] X. Hu and P. Mordohai. "A Quantitative Evaluation of Confidence Measures for Stereo Vision". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 34.11 (2012), pp. 2121–2133 (cit. on pp. 35, 37, 38).

[18] G. J. Iddan and G. Yahav. *Three-dimensional imaging in the studio and elsewhere*. 2001 (cit. on p. 11).

[19] S. Mattoccia. "A locally global approach to stereo correspondence". In: *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. 2009, pp. 1763–1770 (cit. on pp. 2, 3, 41).

[20] S. Meister, R. Nair, and D. Kondermann. "Simulation of Time-of-Flight Sensors using Global Illumination." In: *VMV*. Ed. by M. M. Bronstein, J. Favre, and K. Hormann. Eurographics Association, 2013, pp. 33–40 (cit. on p. 47).

[21] *Mesa Imaging*. URL: `http://www.mesa-imaging.ch` (cit. on pp. 6, 11).

[22] *Microsoft*. URL: `http://www.microsoft.com` (cit. on p. 11).

[23] *Middlebury Stereo Evaluation*. URL: `www.middlebury.edu/stereo` (cit. on pp. 29, 47).

[24] F. Mufti and R. Mahony. "Statistical analysis of signal measurement in time-of-flight cameras". In: *ISPRS Journal of Photogrammetry and Remote Sensing* (2011) (cit. on pp. 12, 16, 22).

[25] *OpenCV*. URL: `http://www.opencv.org` (cit. on p. 32).

[26] L. Pancheri, N. Massari, and D. Stoppa. "SPAD Image Sensor With Analog Counting Pixel for Time-Resolved Fluorescence Detection". In: *Electron Devices, IEEE Transactions on* 60.10 (2013), pp. 3442–3449 (cit. on p. 11).

[27] D. Pfeiffer, S. Gehrig, and N. Schneider. "Exploiting the Power of Stereo Confidences". In: *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*. CVPR '13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 297–304 (cit. on p. 35).

[28] *PMD Technologies*. URL: `http://www.pmdtec.com` (cit. on p. 11).

[29] *Point Grey Research*. URL: `http://www.ptgrey.com/products/stereo.asp` (cit. on p. 27).

[30] M. Reynolds, J. Dobos, L. Peel, T. Weyrich, and G. Brostow. "Capturing Time-of-Flight data with confidence". In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. 2011, pp. 945–952 (cit. on p. 20).

[31] D. Scharstein, R. Szeliski, and R. Zabih. "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms". In: *Stereo and Multi-Baseline Vision, 2001. (SMBV 2001). Proceedings. IEEE Workshop on*. 2001, pp. 131–140 (cit. on pp. 28, 32).

[32] *Softkinetic*. URL: `http://www.softkinetic.com` (cit. on p. 11).

[33] G. Somanath, S. Cohen, B. Price, and C. Kambhamettu. "Stereo+Kinect for High Resolution Stereo Correspondences". In: *Third Joint 3DIM/3DPVT (3DV) Conference*. 2013 (cit. on p. 3).

[34] D. Stoppa and F. Remondino. *TOF Range-Imaging Cameras*. Springer, 2012 (cit. on p. 6).

[35] R. Szeliski. *Computer Vision: Algorithms and Applications*. New York: Springer, 2010 (cit. on p. 9).

[36] C. Tomasi and R. Manduchi. “Bilateral Filtering for Gray and Color Images”. In: *Proceedings of the Sixth International Conference on Computer Vision*. ICCV ’98. Washington, DC, USA: IEEE Computer Society, 1998, pp. 839– (cit. on p. 19).

[37] 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*. Ed. by C. San Martin and S.-W. Kim. Vol. 7042. Springer Berlin Heidelberg, 2011, pp. 214–222 (cit. on p. 14).

[38] A. Wedel, A. Meissner, C. Rabe, U. Franke, and D. Cremers. “Detection and Segmentation of Independently Moving Objects from Dense Scene Flow”. In: *Proceedings of the 7th International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition*. EMMCVPR ’09. Bonn, Germany: Springer-Verlag, 2009, pp. 14–27 (cit. on p. 38).

[39] Q. Yang, K.-H. Tan, W. B. Culbertson, and J. G. Apostolopoulos. “Fusion of active and passive sensors for fast 3D capture.” In: *MMSP*. IEEE, 2010, pp. 69–74 (cit. on p. 2).

[40] P. Zanuttigh, A. Zanella, F. Maguolo, and G. M. Cortelazzo. “Transmission of 3D Scenes over Lossy Channels.” In: *Int. J. Digital Multimedia Broadcasting* 2010 (2010) (cit. on p. 23).

[41] L. Zhang, B. Curless, and S. M. Seitz. “Spacetime Stereo: Shape Recovery for Dynamic Scenes”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Madison, WI, June 2003, pp. 367–374 (cit. on p. 48).

[42] J. Zhu, L. Wang, J. Gao, and R. Yang. “Spatial-Temporal Fusion for High Accuracy Depth Maps Using Dynamic MRFs”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 32.5 (2010), pp. 899–909 (cit. on pp. 2, 59).

[43] J. Zhu, L. Wang, R. Yang, and J. Davis. “Fusion of time-of-flight depth and stereo for high accuracy depth maps”. In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. 2008, pp. 1–8 (cit. on p. 2).