TU Berlin, Fakultät IV, Computer Graphics

Real-time depth imaging

vorgelegt von
Diplom-Mediensystemwissenschaftler
Uwe Hahne
aus Kirchheim unter Teck, Deutschland

Von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften
— Dr.-Ing. —

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr.-Ing. Olaf Hellwich
Berichter: Prof. Dr.-Ing. Marc Alexa
Berichter: Prof. Dr. Andreas Kolb

Tag der wissenschaftlichen Aussprache: 3. Mai 2012

Berlin 2012
D83

For my family.

Abstract

This thesis depicts approaches toward real-time depth sensing. While humans are very good at estimating distances and hence are able to smoothly control vehicles and their own movements, machines often lack the ability to sense their environment in a manner comparable to humans. This discrepancy prevents the automation of certain job steps. We assume that further enhancement of depth sensing technologies might change this fact. We examine to what extend time-of-flight (ToF) cameras are able to provide reliable depth images in real-time. We discuss current issues with existing real-time imaging methods and technologies in detail and present several approaches to enhance real-time depth imaging. We focus on ToF imaging and the utilization of ToF cameras based on the photonic mixer device (PMD) principle. These cameras provide per pixel distance information in real-time. However, the measurement contains several error sources. We present approaches to indicate measurement errors and to determine the reliability of the data from these sensors. If the reliability is known, combining the data with other sensors will become possible. We describe such a combination of ToF and stereo cameras that enables new interactive applications in the field of computer graphics. In addition, we show how the fusion of multiple exposures entails improved measurements and extended applications.

Zusammenfassung

Diese Arbeit beschreibt Lösungsstrategien zur Realisierung bildbasierter Tiefenmessungen in Echtzeit. Während Menschen sehr gut im Schätzen von Entfernungen sind und somit ihre eigenen Bewegungen und auch Fahrzeuge problemlos steuern können, fehlt Maschinen häufig die Fähigkeit, ihre Umgebung in einer dem Menschen vergleichbaren Weise wahrzunehmen. Diese Diskrepanz verhindert es, dass bestimmte Arbeitsschritte automatisiert werden können. Wir gehen davon aus, dass eine Verbesserung der Methoden zur bildbasierten Tiefenmessung einen entscheidenden Schritt darstellt, diese Lücke zu schließen. Wir untersuchen, inwieweit Tiefenbildkameras verlässliche Daten in Echtzeit liefern. Dabei zeigen wir die Probleme existierender Methoden und Technologien auf und stellen verschiedene Ansätze vor, wie die bildbasierte Tiefenmessung in Echtzeit verbessert werden kann. Wir konzentrieren uns auf sogenannte Lichtlaufzeitverfahren und damit im Speziellen den Einsatz einer Tiefenbildkamera, die auf dem Prinzip des Photonen-Misch-Detektors beruht. Derartige Kameras ermöglichen eine pixelbasierte Distanzmessungen in Echtzeit. Allerdings wird die Messung durch mehrere Fehlerquellen beeinflusst. Wir präsentieren Ansätze zur Erkennung dieser Fehler und damit eine Möglichkeit zur Bestimmung der Zuverlässigkeit der Daten dieser Sensoren. Wenn die Zuverlässigkeit bekannt ist, lassen sich die Daten mit denen anderer Sensoren kombinieren. Wir beschreiben eine solche Kombination aus Tiefenbild- und Stereokamera, die neuartige, interaktive Anwendungen auf dem Gebiet der Computergrafik ermöglicht. Darüber hinaus zeigen wir, dass die Fusion von Mehrfachbelichtungen sowohl verbesserte Messungen als auch erweiterte Anwendungsmöglichkeiten mit sich bringt.

Acknowledgments

First and foremost, I wish to thank Prof. Dr.-Ing. Marc Alexa for giving me the possibility to successfully complete this work. His advice has always been very inspiring and a great help during all the time working with him.

I also would like to express my gratitude to Prof. Dr. Andreas Kolb for reviewing this thesis and Prof. Dr.-Ing. Olaf Hellwich for taking the chair of the commission.

I am indebted to Timo Göttel, Dr. Elke Gundel and Mathias Eitz for proofreading and thus bringing this work to the next level.

It was a great pleasure to work in the Computer Graphics department and I thank all colleagues for their great support and making our lab a more than convenient workplace – sometimes even in the absence of sanitary facilities. Namely, I am obliged to Helga Kallan, Mathias Eitz and Kristian Hildebrand among many others for lively discussions and advice. In addition, I thank Prof. Dr. Bernd Bickel, Prof. Dr. Heinz Lemke, Prof. Dr. Oliver Brock, Ronald Richter and Lukas Egger for valuable feedback during the rehearsals of my disputation talk.

This dissertation would not have been possible unless the support of all the graduates who I have given guidance on writing their theses. Especially, I appreciate the hard work by Antoine Mischler, Martin Müllenhaupt and Raul Gigea on dealing with the time-of-flight camera.

I am thankful for being able to use plenty software products that were supremely helpful during the creation process of this thesis. These were primarily TeXnicCenter, MikTeX, Microsoft OneNote, Mendeley (thanks to Paul Föckler), JabRef, TortoiseSVN and PuTTY as well as the web services of LEO, Netspeak, Google Translate, Google Scholar, IEEE Xplore and the ACM digital library. I thank all the developers as I appreciate their great work providing these tools. I also would like to express my gratitude to Martin Profittlich and Alexander Strauch from PMDTec as well as Bernhard Büttgen from MESA Imaging for providing technical details about their products.

In addition, I owe my uncle Siegfried Hahne a debt of gratitude as he strongly motivated the completion of this thesis by mentioning that no one else in the extended family has done this before. Besides my parents and other close relatives who have always stood by me and have given me encouragement, I am truly indebted and thankful for the love of my wife Ilona. If it were not for her and our daughter Ida, I would still be lost in reverie about this thesis.

Contents

Chapter 1

Introduction

This thesis examines to what extend time-of-flight (ToF) cameras are able to provide reliable depth images in real-time and how far these images are utilizable in Computer Graphics (CG) applications. We illustrate how the reliability of such a camera can be enhanced while maintaining the capability of the sensor to capture depth images in real-time.

1.1 Historic development toward machine vision

In early history, machines were invented in order to allow depth estimations in those situations where the human sensing capabilities are not sufficient. A main example is maritime navigation which greatly suffered and still suffers from the sudden appearance of fog making visual navigation impossible. From early ideas like using a bell to indicate the position of ships to the invention of radar, the key interest has always been to allow humans to sense depth when their own capabilities are insufficient.

This paradigm has changed as recent real-time depth sensing devices aim on allowing machines to sense more than just distances. These machines should be provided with ability to gather the complete three-dimensional (3D) scene information including color comparable to how humans sense their environment. In other words, it is no more the goal to create a tool to enhance human depth sensing but to create machines that sense depth by themselves and further process the information. This is usually done in order to support human tasks like manufacturing or navigation. The form of this support can vary from providing human operators with precisely the relevant information they need for solving a certain task to a complete automation of the process. In the latter case, there is no need for any human operator.

There is another interesting aspect. On one side, many solutions exist nowadays to measure distances between large objects like cars or ships in circumstances that would not allow any human to perceive anything. On the other side, sensors exist that are able to detect bumps on surfaces in the range of a few microns that a human could not sense. However, there is no sensor that is a able to sense in both scales. This diversity is reflected in the research on robotics, where still many tasks exist that humans can easily do while machines are far from finding a solution. The robotics research community is somehow close to create a robot that is “as intelligent as a sheep”[1]. However in practice today, most robots are very specialized and hence restricted to a certain task.

One key aspect of this development are missing technologies for real-time depth sensing in various conditions. This thesis presents approaches that enable the implementation of such technologies.

1.2 Applications and goals in Computer Graphics

While the scientific community of machine vision or Computer Vision (CV) follows the described goal of allowing a computer to sense and understand its environment as completely as possible, the general purpose in the realm of CG is to create images.

These images can be entertaining as in games or movies, informative as in data visualizations or communicative as in teleconferencing. Applications following the latter possibility might realize videoconferencing systems like the famous 3D projection of Princess Leia from the first Star Wars movie. A teleconferencing system like this can be seen as one ultimate goal in CG. Recent approaches [48] are getting pretty close to this goal, however only the face of a person is transferred in this example. While the 3D display technology intuitively appears to be the more challenging aspect, in the scope of this thesis we focus on the more fundamental problem of capturing enough reliable 3D data in order to allow a visual 3D representation like the one [49] used in the system of Jones et al. [48]. Thinking further of transferring 3D scene information leads to approaches that extend television to 3D. These so-called 3DTV applications allow the viewers to freely choose their point of view. This for example would allow everyone to follow a football match from the perspective of the goalkeeper or a striker or even the ball itself.

While 3D movies are entering home theaters, for such an immersive football experience as described in the last paragraph, methods and technologies to capture a 3D scene in real-time without intruding the game are still missing. At the time, it is not possible to capture the complete data in order to redisplay it somewhere else. Hence, it is necessary to deal with incomplete data and compute missing elements

[1] Sir Michael Brady at the Queen’s lecture, TU Berlin - 14.12.2011

in order to provide the intended display experience. Here, the computer graphics become an issue. Keeping the football example in mind, it is necessary to produce images from each available perspective. If the perspective can be freely chosen by the viewer at home, the creation of the images has to be done also at home, because it is impossible to broadcast the whole space of perspectives. However, a certain amount of 3D geometric information about the scene is indispensable to produce reliable images. Therefore, fast imaging methods for depth and color are necessary in the process of recreating 3D scenes in real-time. This thesis presents approaches toward this goal and demonstrates among other things an exemplary application for Augmented Reality (AR).

AR is a technology that combines real world video footage with virtual content. If the virtual content is to be seamlessly integrated, it will be necessary to know the geometry of the captured scene. Knowing the geometry enables real world objects to occlude virtual content. Besides this, it becomes possible to calculate realistic illuminations for the virtual objects including the cast of shadows onto real world objects. Shadows and occlusion are one of the most important monocular depth sensing criteria of the human visual system. Hence, their absence would strongly disturb the immersion. If the virtual content is meant to be embedded in live video, it will be further indispensable that the geometry information is available in real-time. Real-time depth imaging provides this knowledge about the geometry and we show that the methods developed for this thesis are capable of being used in AR settings.

1.3 Reliability is the key

Before presenting a broad variation of methods that allow depth sensing in real-time, we briefly want to discuss the most important point of all these systems: *reliability*. We therefore depict examples from various fields of application.

Maritime navigation First, reliability is crucial if the sensor is used in the traditional form as a tool to replace the human depth sensing abilities. In maritime navigation, one has to be sure that there is no other ship when navigating through a narrow passage in the presence of fog. In such a case, a failing sensor can even harm people.

Automotive applications If we transfer this application to the automotive area, reliability is still crucial but in case of a driver assistance system, it is not the last instance. Recent systems usually avoid collisions only if the car is moving with less than 30 km/h[2]. If the car is driving faster and a sensor detects an issue, the

[2] http://www.bosch-kraftfahrzeugtechnik.de/media/db_application/downloads/pdf/safety_1/de_6/Vorausschauendes-Notbremssystem.pdf

system will first warn the driver and then support his actions, while the responsibility still lies in the hands of the driver. This will change if we think about autonomous vehicles or robots that do not need any driver or operator. By now, in general, robots are not allowed to operate in the same area as humans do. This security principle is standard due to the missing reliability of existing sensors. There is no absolute evidence that the robot operates correctly and as it might harm people nearby, such a setting is not permitted. Note that the reliability of humans is limited as well, which makes the issue even more complex. In the remainder of this thesis, we will consistently come back to examples from the automotive area. On one hand, such examples are comprehensible due to the high occurrence in every day situations. On the other hand, nowadays depth sensing devices are already embedded in cars.

Gaming Another more recent field of application for depth sensing is the consumer gaming market. Here, we can observe a trend away from classic input devices like the gamepad toward more natural interfaces. These interfaces can be small hand-held devices with inertia and acceleration sensors or even the user can be the interface itself. Most prominently, Microsoft claims to promise "a gaming experience that is safe, secure and fun for everyone" with their Kinect sensor[3]. At first sight, the reliability of the system in case of a game seams less crucial. However, due to the permanent and direct interaction between the user and the device, any kind of error, such as delay, imprecision or interruptions, quickly creates a great deal of annoyance that potentially discourages and frustrates the user. Finally, this could avoid acceptance of such a device as a reliable gaming interface.

Following these examples, we argue that reliability is the key for real-time depth sensing. Hence it is always the goal in developing real-time depth sensors to reach an as high as possible degree of reliance. This thesis presents approaches toward this goal for ToF imaging.

In the realm of CG, we see this depth sensing technology as the most promising. It provides depth and intensity data at very high frame rates. The compact setup allows the combination with other cameras and sensors which is the second main aspect of this thesis. Combining several depth imaging principles allows us to even out the drawbacks of each principle. We demonstrate how systematic errors can be reduced at low computational costs so that the real-time ability of the ToF imaging devices is kept. In addition, our solutions are low cost and avoid any special hardware or elaborate training periods. This makes them quickly applicable in practice and reproducible. We claim that our approaches are an important step in the development of real-time depth sensing toward utilizing machines with depth sensing abilities that are comparable to humans.

[3] `http://www.xbox.com/en-US/kinect(accessed:03.02.2012)`

1.4 Outlook

In order to reconstruct the previously described development from a human tool to machine vision, the most promising depth sensing methods are described in the following Chapter 2. As an introduction, we explore how humans sense depth in order to understand some of the methods as they follow examples from nature.

In the subsequent chapter, we first introduce ToF imaging in general and discuss the issues that reduce the reliability of such sensors. In addition, we present related work about reducing the errors of ToF cameras. Existing ToF cameras lack reliability since they have strong constraints about the environments they are utilized in. These depth sensors have no ability to capture color information which is necessary for most applications in the field of CG. Their flexibility is further narrowed as they have a reduced dynamic range.

Chapter 4 describes a proof-of-concept that the reliability of ToF imaging system can be enhanced by combining these sensors with stereo depth sensing. We combine a low resolution ToF depth image camera based on the photonic mixer device (PMD) principle with two standard cameras in a stereo configuration. We show that this approach is useful even without accurate calibration. In a graph cut approach, we use depth information from the low resolution ToF camera to initialize the domain, and color information for accurate depth discontinuities in the high resolution depth image. The system is promising as it is low cost, and naturally extends to the setting of dynamic scenes, providing high frame rates. This chapter has been published as [36] and [37].

In Chapter 5, a system is presented that implements such a combining approach. Besides enhancing the reliability of the data, it enables capturing color information in real-time. We show some exemplary applications from the realm of AR. We present a framework for computing depth images at interactive rates. Our approach is based on combining ToF range data with stereo vision. We use a per-frame confidence map extracted from the ToF sensor data in two ways for improving the disparity estimation in the stereo part. First, we use the map together with the ToF range data for initializing and constraining the disparity range. Second, the map together with the color image information allows us to segment the data into depth continuous areas, enabling the use of adaptive windows for the disparity search. The resulting depth images are more accurate than from either of the sensors. In an example application, we use the depth map to initialize the z-buffer so that virtual objects can be occluded by real objects in an AR scenario. This chapter has been published as [38].

In Chapter 6, the concept of confidence is further evaluated. We present the adaptation of a method known from computational photography that allows us to enhance the reliability and the dynamic range of a ToF camera in real-time without the need of any calibration procedures in advance. This chapter deals with the problem of automatically choosing the correct exposure (or integration) time for ToF depth image capturing. We apply methods known from high dynamic range

(HDR) imaging to combine depth images taken with differing integration times in order to produce high quality depth maps. We evaluate the quality of these depth maps by comparing the performance in reconstruction of planar textured patches and in the 3D reconstruction of an indoor scene. Our solution is fast enough to capture the images at interactive frame rates and also flexible to deal with any amount of exposures. This chapter has been published as [39].

In a concluding chapter, we summarize and discuss the presented approaches and close the circle to our described development of real-time depth sensing from a human tool to machine vision.

Chapter 2

Depth imaging

This chapter will be an overview about existing approaches to measure distances between a sensor and solid objects with the aid of a computer. We focus on methods that provide not only single point distances but complete images. We only deal with systems that ensure that the measurements are completed in real-time. This means that the method should allow the capture of dynamic scenes. The terms *real-time* and *dynamic* have to be handled with care. In the scope of computer graphics and thereby this thesis, the universal goal is to produce images for human inspection. Hence, the necessary dynamic temporal range is limited by the human visual system. Humans are not able to distinguish between a fast sequence of static images and real motion. We can exploit this fact – which is the foundation for cinema – and call any imaging system that is able to produce at least 24 images per second a *real-time imaging system*.

In order to let the reader receive an impression on the complexity of the depth imaging problem, we first give a brief introduction how humans perceive distance information. A core contribution of this thesis is the depiction that the combination of several technologies leads to an improvement of depth sensing devices. We show that the human visual system is combining several depth imaging methodologies, too. After that, we define the difference between imagery and imaging. Then, we compare imaging methods according to their usability in various applications before we go into details about the most promising methods.

2.1 Human abilities for real-time depth sensing

Vision The eyes are the first part of the human visual system. They collect light and project it onto the retina from where it is further processed to the visual cortex. The basic information that can be extracted from the incoming light rays is color information. Visual depth sensing is therefore based on the processing of the received color information.

Depth sensing is a learned ability and humans sense depth primarily by vision. While the two human eyes allow stereopsis – the fusion of both images, then extracting disparities and hence the reconstruction of the distance of any object – the imaging quality is surprisingly poor. Helmholtz detached that the technical properties of the eye are far from what is perceived by humans. He deduced that many effects come from unconscious inference [119].

In order to reconstruct depth from two stereo images, the human brain, or more precisely the visual cortex, has to find corresponding features in both retina images. Due to geometric properties, these features lie on a line through both images – the so called epipolar line. This epipolar geometry has been used since many years [21] to reduce the computation costs for stereo algorithms. Rectification of the two stereo images restricts the search range for corresponding features to a single line. Quite recently, Schreiber et al. [105] has found that humans move their eyes in order to reduce the computational load for the brain. In other words, humans tend to rectify their eyes which reduces the complexity of finding correspondences by one dimension.

Besides Helmholtz' explanation for human depth sensing, there is the ecological approach which has been introduced by Gibson [29]. He claims that human depth perception is strongly affected by the motion of a person. The optical flow of feature points gives an impression of the position and direction of movement of a person. In addition the surrounding environment is geometrically interpreted based on texture gradients.

Similarly to stereopsis, these two principles of depth perception have both been used in computer vision to extract depth information from images. First, so-called structure from motion (SfM) approaches compute the optical flow from a sequence of images and reconstruct the 3D information of the scene. Second, texture gradients have also been used to extract longitudinal surfaces from images [3].

Another important aspect for humans to judge the distance of objects correctly are the so-called *monocular depth criteria*. These criteria include occlusion, perspective, relative size and height, but also atmospheric perspective, expected size, shadows as well as the already mentioned texture gradients. While we refer to the book by Goldstein [30] for a complete description of the human depth sensing abilities, it is worth mentioning that humans estimate distances always according to intended actions and the effort of these actions which is linked to the distance. Hence, in contrast to machines, the visual human depth sensing does not only depend on environmental influences but also on the personal condition.

Sound Sound is an alternative as vision is not given in all situations. Blind people orientate from the surrounding noise (passive) or even try to use human echolocation [117] (active).

After Goldstein [30], the human auditive system uses three principles in order to localize an acoustic source:

1. Head-related transfer function (HRTF)

2. Interaural time difference (ITD)

3. Interaural level difference (ILD)

The HRTF describes how the sound signal is transferred to a spectral stimulus in one ear and allows humans to sense the direction of a sound. Besides this monoaural stimulus humans use both the difference in time as well as the difference in intensity between their left and right ear to localize an acoustic source.

Here we can extract technical principles. The acoustic localization uses triangulation as the position is determined from the known geometry of the human head. In addition, the HRTF can be compared to photometric approaches like shape from shading (SfS). As the sound is perceived differently depending on its source position, humans are able to localize the direction.

These principles are applied in several technologies that we describe in the next section.

2.2 Real-time depth imaging methods

This section introduces several methods for real-time depth imaging. First, the term *depth* or *range imaging* is defined and then several methods are discussed. We classify the methods by their optical principles. This classification aims to provide the reader with a general understanding of various depth sensing methods. This understanding is necessary in order to be able to comprehend the core contributions of this thesis.

Preliminaries The MacMillan[1] thesaurus defines *imaging* as "the process of producing an image by using a machine that passes an electronic beam over something", while the mathematical definition of an image from Wolfram Mathworld[2] is as follows:

"If $f : D \to Y$ is a map (a.k.a. function, transformation, etc.) over a domain D, then the image of f, also called the range of D under f, is defined as the set of all values that f can take as its argument varies over D, i.e.,

$$\text{Range}(f) = f(D) = \{f(X) : X \in D\}.$$

'Image' is a synonym for 'range', but 'image' is the term preferred in formal mathematical writing."

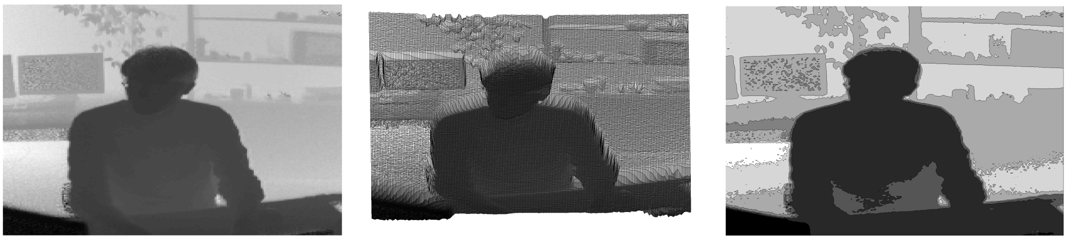
[1] http://www.macmillandictionary.com/thesaurus/american/imaging

[2] http://mathworld.wolfram.com/Image.html

These contradicting definitions might lead to some confusion because the term *range imaging* is very common in the computer vision community. We further distinguish between *imaging* and the term *imagery*, which indicates a set of images, while imaging describes the process to produce this set.

Besl [6] defines "a range-imaging sensor [as] any combination of hardware and software capable of producing a range image of a real-world scene under appropriate operating conditions. A range image is a large collection of distance measurements from a known reference coordinate system to surface points on object(s) in a scene." In this thesis, we use this terminology. In addition, we use *depth image* and *depth map* as synonyms for *range image*. The term *distance* is used for distances between two points not images.

Representation Note that Besl also lists a number of synonyms for range image and illustrates common ways to display those images. A range image is usually a matrix of numbers just as a conventional digital intensity image. Hence, the natural representation is a gray-level image where the intensity does not refer to the irradiance or brightness as in conventional images but to the distance of an object point to the sensor's center. Additionally, this data can also be displayed as a surface plot which might provide a better 3D impression to the viewer. Besides this, there are contour maps that are mainly used in cartographic applications and therefore outside the scope of this thesis. Figure 2.2 shows different representations of the same scene produced by the MATLAB functions **imagesc, surf** and **contourf**. Note that the intensity is always proportionally mapped to the distance, that is why objects further away are brighter than close ones.



(a) Gray-level intensity

(b) Surface plot

(c) Contour plot

Figure 2.1: Exemplary representations of a depth image.

Performance measure We follow the classification of Besl in six different optical principles: *radar, active triangulation, Moiré, holographic interferometry, focusing and Fresnel diffraction*. Additionally, we include a seventh category named *passive triangulation* that includes stereo imaging which has become a well-studied topic during the last decades and is applied in our approach. In order to compare recent existing approaches that are based on different principles in terms of performance, we define – as proposed by Besl – the merit of performance

$$M = \frac{L_r}{\sigma_r \sqrt{T}}$$

where L_r is the depth of field, σ_r is the root mean squared (RMS) range accuracy and T is the pixel dwell time – the time required for a single pixel measurement. This merit allows the comparison of sensors independent from their underlying principles. In this thesis, we compute the merit for chosen recent systems in order to give an impression on the development of the methods, while Besl [6] gives a rather complete overview about existing systems in 1988. Such a complete survey is not the aim of this thesis, therefore only some relevant systems are described in the following.

Radar The radar principle is simple: send out a signal and measure the time until the reflection arrives. This principle is used by animals like bats [32] and porpoises [56], but also by blind humans [117]. In 1904, Christian Hülsmeyer demonstrated a first working prototype that detected a ship in over 100 meters distance [113]. After this initial proof of concept the technology was further developed until it became extensively used in World War II. Nowadays every ship and airplane has to be equipped with a radar sensor. Additionally, most new cars use radar technology to sense distances between the car and surrounding objects. This enables the car to e.g. support parking or to trigger emergency brakes.

The term *radar* is generally used for depth sensing methods that emit and receive any kind of radio waves – radar stands for *radio detection and ranging*. As radio waves are transmitted with speed of light, the challenge is to measure the time between emission and reception as precisely as possible. We call this the *time-of-flight (ToF)*. This can be realized by modulating the signal. A modulation changes the characteristics of a signal so that information can be embedded into the signal. If a periodic sinusoidal carrier signal is used, either frequency, amplitude or phase of this signal can be modulated in order to encode information. Measuring these characteristics is simpler and allows time measurements in a range that is adapted to the application.

In the field of machine vision, there are many approaches using a specialization of radar called light detection and ranging (LIDAR). In this case, light is emitted and received by photo receptors. Again, the light has to be modulated as the ToF is hardly measurable directly. Note that there are recent technologies that allow such measurements in laboratory environments [84].

There is a dichotomy between scanning and scannerless technologies. In scanning technologies usually a laser beam sweeps over the target and thereby *scans* the object. More recently, infrared (IR) light emitting diodes (LED) have been used to avoid this scanning procedure as it involves movable parts that are prone to errors from drift or other mechanical issues. Note that the term laser scanner is also common even if a different optical principle namely triangulation is used.

As one of the first LIDAR devices, in 1977 Nitzan et al. [87] presented a one-dimensional (1D) scanner that uses a single laser beam and a scanning mirror that controls the direction of the laser beam. The scanner produces 128 × 128 pixel images at a range of 1 – 5 m while the resolution is about 1 cm. It usually takes up to 2 hours for one image. The exact time depends on the reflection properties of the scene as the integration time is adapted per pixel. The system reaches a merit of $M_{Nitzan} = 3770$.

Nitzan et al. identify that the dominant noise is photon noise due to bad reflections. They conclude that the high dynamic range of such systems is the key problem with ToF sensing. The dynamic range can exceed 100 dB as there is a large difference between close and bright (or highly reflective) objects and objects located far away from the sensor as well as dark ones. Nitzal et al. calibrate the system using "a standard white-mat sample made by spraying 20 thin layers of barium sulfate on an aluminum substrate" whose reflection properties at the wavelength of the laser are known. This allows to make assumption on the error of the system and therefore gives an intention on the reliability of the measurements.

Following Besl [6], there are many possible variations of the radar principle that depend on the modulation of the emitted signal. Two prominent variants for LIDAR systems are pulse-code modulation and amplitude modulation. All the systems Besl described in 1988, suffer from the necessity to scan the scene and therefore, the maximum possible frame rate is limited. Not until the late 1990ies this limitation vanished, when for both modulation schemes a chip has been developed that allows a parallelization.

On one hand, we have so called ToF-cameras based on the PMD principle [108, 83, 124]. Here, the reference signal is an amplitude modulated near-infrared (NIR) light front that illuminates a whole scene. The reflections are captured by a chip that is able to correlate the received signal with the reference signal per pixel. By now, this allows a parallel measurement of about 200 × 200 depth values. The chip is based on complementary metal-oxide-semiconductor (CMOS) architecture and is therefore mounted as a regular camera using standard optics resulting in a field-of-view of 40° by 40°. For the commercially available PMD[vision] CamCube 3.0 camera, the merit of performance is $M_{PMD} \approx 2\,825\,000$.

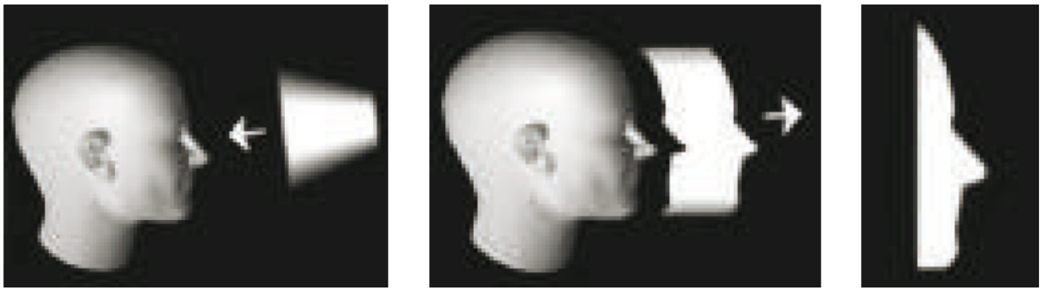


Figure 2.2: Nano-shutter camera operation principle (extracted from [35, 92]).

On the other hand, there are ToF cameras using a so-called nano-shutter that

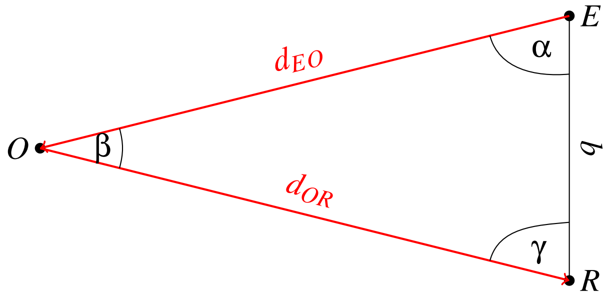


Figure 2.3: The triangle between object point O, emitter E and receiver R.

has been developed by Iddan and Yahav [47]. It allows to cut a light front so that the captured intensity directly relates to the distance of the object. This process is illustrated in Figure 2.2. Only few research has been done about the precision of this sensor. While Gvili et al. [35] show that the sensor is utilizable in applications like depth keying, Radmer and Krüger [92] present absolute distance measurements with an offset of more than 20%.

As already predicted by Nitzan [87] in 1977, the ToF principle is very promising and so many manufacturer provide commercially available devices. Recently, Piatti [91] gives an overview and comparison of available devices. The principle of measuring direct reflections allows very compact form factors that even allow endoscopic applications [90]. This is the main advantage of this methods against active triangulation.

Active triangulation Simply said, triangulation is realized by setting up a triangle between object O, receiver R and emitter E. The distance d_{OR} between object and receiver can be reconstructed, if one side and two angles of the triangle are known. Usually, the distance between receiver and emitter is known as the baseline b which forms one side of the triangle. From both the emitter and the receiver the direction of the signal toward the object has to be known. These directions hold the adjacent angles α and γ. We derive

$$d_{OR} = \frac{b \sin \alpha}{\sin \beta}$$

where β is the opposing angle to the baseline b as depicted in Figure 2.3. Its value derives from the fact that all three inner angles of a triangle sum up to 180°.

We conclude that any depth sensing method based on triangulation needs to be able to determine three parameters: the baseline, the direction of the received signal and the direction of the emitted signal.

The classic active triangulation device is a laser scanner. It is built up of a laser diode the emits a light beam and some receiver that is able to detect the reflection of the light beam. The baseline is formed by these two units and each of them has to be able to determine the direction of the signal which is the light ray in this case.

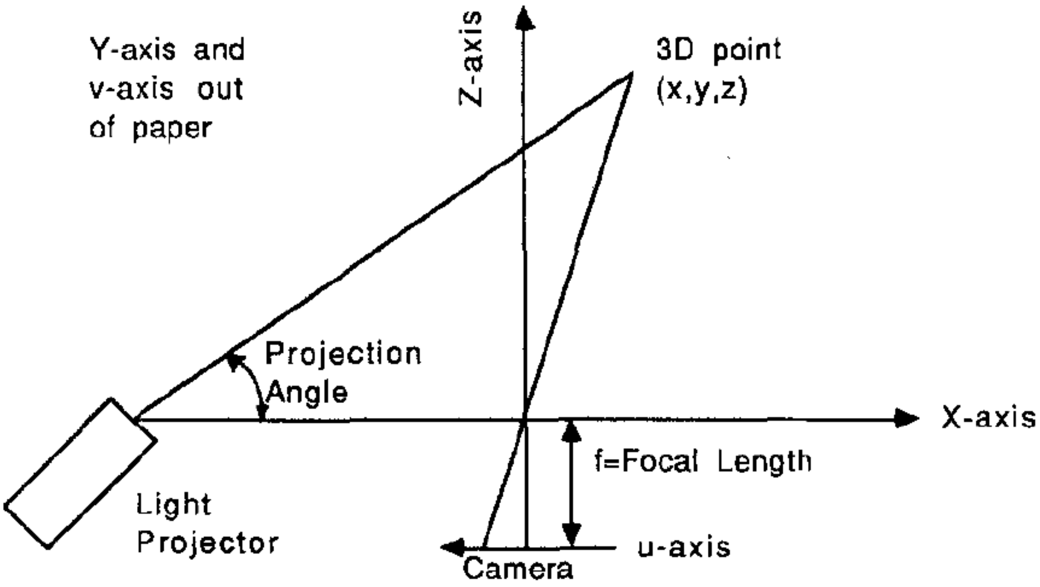


Figure 2.4: Camera-centered active triangulation geometry from [6].

The laser is therefore either mounted on a sweeping device or more commonly in front of a tilting mirror in order to scan the scene. As receiver a camera can be used that is sensitive to the wavelength of the laser beam. This sensitivity allows a fast detection of the reflection that is necessary to determine the direction.

Figure 2.4 illustrates how the measurement is realized in practice. The 3D coordinates of a 3D point $O = (x,y,z)^T$ can be computed as

$$O = \frac{b}{f \cot\theta - u} I$$

where f is the focal length of the camera, θ is the projection angle of the emitted beam and I is a vector $(u,v,f)^T$ containing the pixel coordinates u and v as well as the focal length.

This process can be further enhanced by emitting a plane of light instead of a single ray. This produces a deformed line in the receiver image from which the distance can be calculated.

Such a system allows a very high precision and has been successfully applied for cultural heritage in the Digital Michelangelo project [67]. Here, the focus lies on high precision. The acquisition time does not influence the quality as only static scenes like the David statue have been scanned. Nevertheless, there are efforts to enhance the acquisition speed of laser scanning devices. The key drawback are the moving parts that are necessary to sweep the emitted signal and scan the object. This makes these devices either imprecise or very costly.

At the one end of the spectrum there is the DAVID 3D scanner[3]. Here, the sweeping of the laser has to be done manually by a human operator. This makes it not applicable in industrial applications as the scanning process is not repeatable. At the other end, we have professional scanning products like the FARO Focus 3D[4] that captures 976 000 measurement points per second or the Velodyne HDL-32E[5] with 700 000 3D points per second.

For the Faro Focus 3D, this results in a merit of $M_{Faro} = 12\,349\,000$. However, even products with such a high merit do not necessarily provide real-time depth images. The scanning process for the FARO Focus 3D – with an field of view of 360° in horizontal and 300° vertical direction – takes a couple of minutes. If we assume real-time – as in conventional video – needs at least 24 frames per second (fps), we would have to reduce the field of view of the laser scanner to less than 2° in both (horizontal and vertical) directions.

In order to capture dynamic scenes in 3D the process of triangulation can be highly parallelized. All existing approaches have already been presented by Besl [6] in 1988. While active triangulation methods can generally also be called structured light methods, in recent literature the term *structured light* is mainly used for approaches beyond point and line. Besl names five further categories – *Miscellaneous, Coded Binary Patterns, Color Coded Stripes, Intensity Ratio Sensor* and *Random Texture*.

The last method of random textures has recently become very popular through the Microsoft Kinect device. This sensor emits a random dot pattern in the NIR range and captures this scene with a camera. From the deformation of the pattern, the depth information is reconstructed. The system captures depth images at a resolution of 640 × 480 pixels in a range of 0.5 to 5 m at 30 fps. We follow the analysis of Khoshelham [57] and assume a precision of 4 cm at the maximal distance – the error increases quadratically with the distance. This results in a merit of performance $M_{Kinect} = 341\,530$.

As there is no official detailed description about this algorithm there are plenty highly speculative explanations around. Victor Castaneda and Nassir Navab give a good overview in their lab course slides[6], however it is based on assumptions and unreferenced images from patents. There are three patents [26, 127, 25] by PrimeSense that explain the principle. The authors have further published two articles [27, 81] that describe the novelty. The pattern of the projection depends on the distance. For each distance the pattern has to be determined once in a preprocessing step. Then the distance is computed by a cross-correlation of image parts with the predefined distance patterns. The depth resolution in this approach is limited on one hand by the spatial resolution of the camera that captures the projection of the pattern and on the other hand by the computation of the cross-

[3] www.david-laserscanner.com

[4] www.faro.com/focus/

[5] http://velodynelidar.com/lidar/hdlproducts/hdl32e.aspx

[6] http://campar.in.tum.de/twiki/pub/Chair/TeachingSs11Kinect/2011-DSensors_LabCourse_Kinect.pdf

correlation. For the Kinect sensor, this approach has been realized for only three different distance levels so-called reference planes while the precise distance is determined by the shift of the speckle pattern as illustrated by Khoshelham [57] and in Figure 2.5. The unknown distance Z_k can be computed by

$$Z_k = \frac{Z_o}{1 + \frac{Z_o}{f\,b} d}$$

where d is the observed disparity in image space. The reference distance Z_o, the focal length f and the baseline b have to be determined by calibration.

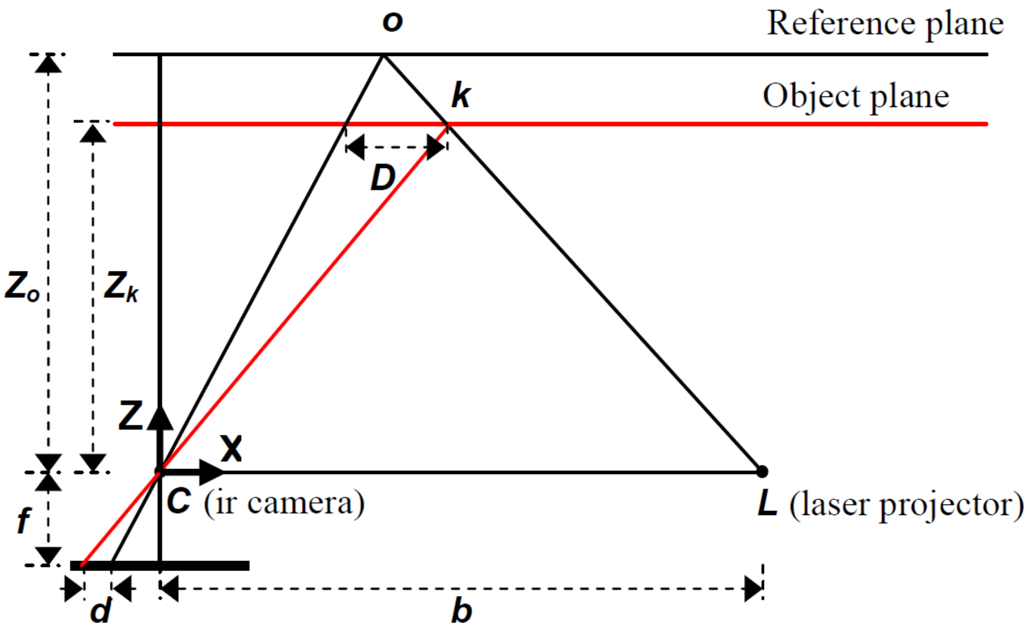


Figure 2.5: Active triangulation and disparity with reference plane from [57].

As a first observation, we can maintain that it is promising to combine several basic principles like in this case the active triangulation with speckle interferometry. We will discuss this aspect further in the remainder of this section.

Passive triangulation Note that so far we have only described active triangulation. If the emitter is replaced by a second receiver, a passive stereo setup is created. Passive stereo is one of the humans depth sensing methods as described in Section 2.1. Therefore it is intuitively a very promising method toward a humanoid machine vision. Thus, we briefly discuss stereo imaging in the following, although it is not included in the survey of Besl [6]. For a complete mathematical description we refer to the books of Faugeras [16] as well as Hartley and Zisserman [40].

Passive stereo imaging strongly suffers from its high computational load. In order to calculate the triangulation, corresponding points or features have to be matched in both images. The offset between the points in the two images is called disparity. From the disparity the distance can be calculated if the intrinsic and extrinsic parameters of both receivers – usually cameras – are known. We refer to the paper of Zhang [128] for a complete description of the camera parameters.

While the problem of finding corresponding points can be reduced to a 1D problem by rectification of the camera images, it remains ill-posed. Due to occlusions, some parts of the one image are not visible in the other and hence no depth information can be extracted from the disparity for these parts. This makes the matching problem difficult as it introduces features that do not have a matching counterpart.

Most of the existing matching algorithms define a cost function that returns a probability that two pixels from different camera images refer to the same real-world object point. These cost functions compute a so called feature that is defined on a small area around the image point as a single pixel does not hold enough information. The larger this area the higher the computational load. However, the reliability grows with the area as long as depth homogeneity is given – this is the case if the area only covers the same real-world object. The determination of this coverage is essential for adapting the size of the area. A solution for this problem is presented in Chapter 5.

Usually, the cost function returns values according to the similarity of the features, therefore both cameras should provide similar images in terms of sharpness, brightness and contrast. To achieve this the field of view should overlap as much as possible. However, a longer baseline allows more accurate results because the determination of angles in the triangle is limited by the spatial resolution of the cameras. Alternatively, it is possible to compute features that are independent of some of the image properties, however their computation is costly again. This issue is discussed in detail by Szeliski [110].

A further issue with stereo correspondence is specularity. Specular reflections are viewpoint dependent, hence specular highlights in the left and right stereo image occur on slightly different positions. Unfortunately, specular highlights represent a good feature and look similar in both images. Therefore, there are approaches to preprocess the images and detect or even remove the specular highlights [68].

For a complete overview about stereo algorithms and their performance, we refer to the Middlebury website[7] that provides test data in order to make all the existing approaches comparable. The website offers a benchmark and gives an overview about the performance of most of the existing algorithms. A complete description of their methods is given by Scharstein and Szeliski [100].

[7] http://vision.middlebury.edu/stereo/

While already Nitzan et al. [87] claims that the future lies in triangulation and ToF, we discuss the remaining four optical principles from Besl [6] only very briefly in the following.

Moiré Two diffraction gratings are used to project two patterns onto a surface. The relative depth can be extracted from the phase difference of the projected overlaying patterns. We refer to Bartl et al. [4] for a more recent review of this method. It allows the detection of surface roughness in the range of microns.

Fresnel diffraction Similar to Moiré this method is used to measure surface roughness. It is based on the Talbot effect [111] for diffraction gratings which is a natural consequence of Fresnel diffraction. This effect leads to repeating patterns of diffracting light at a distance that depends on the wavelength of the light and the period of the grating [118].

Holographic interferometry This method is also very similar to Moiré, but here, holographic fringe patterns are used instead of diffraction gratings. This results in an even smaller depth of field.

As all of the last three methods are based on special optical effects and are applied in the field of surface analysis, we refer to the book of Kreis [63] for details. However, as already mentioned in the paragraph on active triangulation, these concepts are successfully included in recent real-time technologies like the Microsoft Kinect sensor.

Focusing Knowing the focal length of the camera allows to reconstruct depth from the amount of how much an object is in focus. While the principle – often also called *depth from focus* – has been evaluated by Grossmann [33], Nayar et al. [85] present a system for real-time depth imaging. While their system reaches a merit of $M_{Focus} = 1\,357\,600$, it is also a combination with other approaches as it includes the active projection of a pattern. They use this approach to overcome the problem of the need for high-frequent textures that every passive stereo system has. They use a video projector including a lens and hence a focal length. The projected pattern is in focus only a one certain distance. From the amount of defocus the distance of the object can be obtained. This concept has also influenced a combining approach by Jones and Lamb [50]. They manipulate a camera by creating two apertures. These apertures lead to the projection of two stereo images onto a single image plane. The disparity information has to be extracted from a single image which introduces new issues e.g. the separation of the images.

2.3 Conclusions from depth imaging methods

The general depth sensing methods presented in the last section are clear in their general approach. The radar principle allows a broad range of applications. However it is not trivial to tune a radar setup to the specific conditions where it will be used. In the last two decades, the invention of PMD allows real-time depth imaging of scenes that are comparable with human sensing capabilities. This leads to the possibility to capture scenes in a similar way as in standard photography. However, it is technically challenging to produce depth imagery that is reliable and contains only a limited amount of noise. The noise and limitations in the production lead to systematic errors for PMD based ToF imaging that will be described in detail in the following chapter.

In contrast, triangulation methods suffer from the necessity of the triangle as illustrated in Figure 2.3 on page 13 in the last section. This concept will always lead to missing information about those parts of a scene that are not visible – or not reachable in case of active illumination – by one of the receivers due to occlusions. Occlusion handling is therefore necessary for all real-time depth sensing methods based on triangulation. In addition, occlusion handling is not trivial and therefore introduces higher computation costs that reduce the speed of acquisition which is crucial for real-time depth imaging.

As mentioned, it is promising to combine these general concepts as all methods based on a single concept suffer from systematic errors that come with the concept. This thesis presents solutions to overcome these problems by combining concepts from different fields. All these solutions are predicated on the utilization of ToF cameras based on the PMD principle – also called PMD cameras – that provide erroneous depth images in real-time. We combine these cameras either with other hardware but also software concepts that allow an elimination or at least reduction of the errors.

In the following chapter, we explain the basic working principle in detail and present a set of assumptions that all lead to various errors in the captured data. These insights have to be kept in mind when the core contributions of this thesis toward real-time depth imaging are presented in the remainder.

Chapter 3

Time-of-Flight Imaging

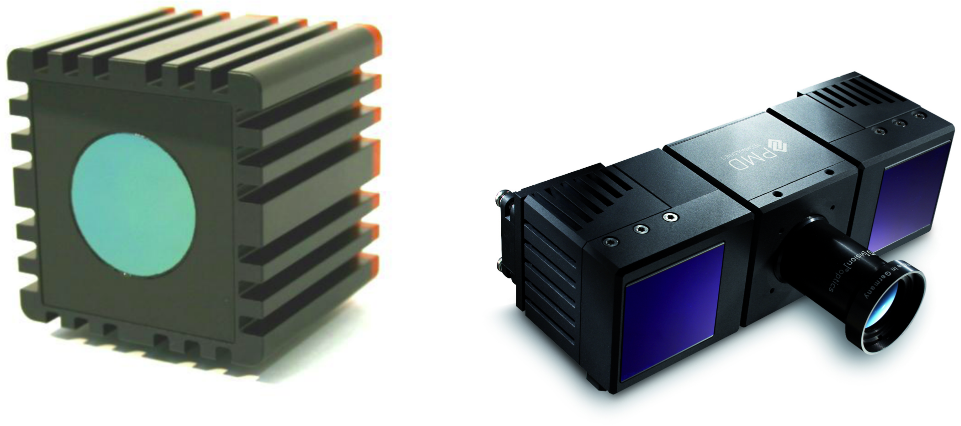
There are several manufacturers that sell PMD based ToF cameras. For a recent overview, we refer to the already mentioned PhD thesis by Piatti [91] and the corresponding Wikipedia article [123]. Since these cameras are commercially available, they are being used in several research groups. Most commonly, research projects use the products of PMDTec/ifm electronics[1] and MESA Imaging[2]. Figure 3.1 displays two exemplary cameras from their product line. While PMDTec evolved from research at the University of Siegen, MESA Imaging is a spin-off from the ETH Zurich. They both provide cameras that are based on the PMD principle and they are the first who put their products on the consumer market. Another company called Canesta has been bought by Microsoft in the process of emerging technologies for *Project Natal* which evolved in the already mentioned Kinect sensor. Canesta offered also ToF cameras that are based on the same basic principles but there is no more contact information available.

Note that the term *time-of-flight camera* is used for all kind of cameras based on the ToF principle. We differ between amplitude modulation, pulse-code modulation and nano-shutter approaches. The cameras used in the experiments presented in this thesis are based on the PMD principle which belongs to the amplitude modulation and works with uncorrelated light from sources like LEDs. This is explained elaborately in the following section.

When seeing a live demonstration of the depth image produced by the sensors for the first time, most researchers are impressed. However, they immediately remark the strong noise in the data. The data is not constant over time for static scenes. If the depth is intensity or color coded this effect is clearly visible. If the depth data is rendered in 3D, flat surfaces look more like troubled waters than wall or table surfaces as illustrated in Figure 3.2. When capturing a human head, the noise makes it hard to distinguish even prominent features as the nose and eyes. Images of surfaces textured with a checkerboard pattern disclose that the distance measurement depends on the reflectivity of the captured objects. We will discuss

[1] `www.pmdtec.com`

[2] `www.mesa-imaging.ch`



(a) Mesa Imaging SR 4000

(b) PMDTec PMD[vision] CamCube 3.0

Figure 3.1: Two exemplary PMD based ToF cameras.

this behavior – illustrated in Figure 3.7 on page 34 – in Section 3.2.

From the trends in the publications at one of the first workshops about PMD based ToF imaging[3], we assume that all these observations lead to the fact that the first step in many research groups was thinking about how to reduce this error. The data is that noisy that existing methods and applications from the field of computer graphics are not applicable directly. The captured depth data has to be processed in advance.

In the following, the basic working principle of the ToF cameras based on the PMD concept is explained in detail. After that, the known error sources are explained and existing literature that aims on the correction of these errors is revisited and discussed.

3.1 Basic principles

As pointed out in Section 2.2, the ToF imaging principle is basically radar. The device emits a signal and receives its reflection. For the cameras used in the scope of this thesis, the emitted signal is an amplitude modulated light signal in the NIR range. The light is reflected by the objects in a scene and these reflections are captured by a sensor. The sensor measures not just the amount of incoming light but also the phase of the signal. As the phase of the emitted signal is known, the phase shift between emitted and received signal can be measured. Knowing the frequency of the signal allows to reconstruct the time the signal needed to travel from the emitter to the object and back to the receiver. When this *time-of-flight* t_d

[3] Dynamic 3D Imaging Workshop in Conjunction with DAGM, September 11th 2007, Heidelberg, Germany, `http://www.zess.uni-siegen.de/pmd-home/dyn3d/workshops/2007/`

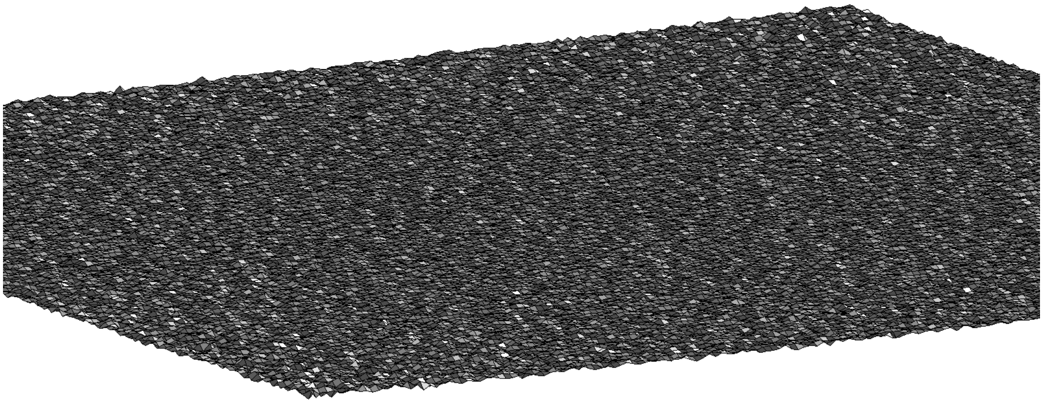


Figure 3.2: Depth data of a wall rendered in 3D.

is known, the distance d can theoretically be calculated as

$$d = \frac{c t_d}{2}$$

where c denotes the speed of light. As it is challenging to measure the time-of-flight t_d directly, the modulation frequency f is used for a substitution. The phase delay φ can be measured and is directly proportional to the time-of-flight t_d. With

$$t_d = \frac{\varphi}{2\pi f} ,$$

the distance can be computed as

$$d = \frac{c\varphi}{4\pi f} .$$

Hence, the challenge is simplified to measure the phase delay φ of the reflected signal. This is done on a special chip called PMD. Details about the chip design is given by Lange [65]. In the following, this measuring procedure is explained in detail.

In literature, there are several variants describing how to compute the depth information from the phase delay of the signal. These descriptions range from very simple but incorrect to highly accurate but barely comprehensible. While Kolb et al. [61] provide a clear but compact and hence simplified explanation, Frank et al. [22] give a precise definition that emphasizes the mathematical properties. In the scope of this thesis, we follow these two references, as we want to provide a clear and precise explanation that allows the reader to comprehend our approaches on enhancing the real-time imaging properties of ToF cameras.

Derivation The emitted reference signal $R(t)$ and the reflected optical signal $S(t)$ are correlated on the camera's chip. Both signals have the same modulation frequency f that leads to the same angular frequency

$$\omega = 2\pi f.$$

As described by Frank et al. [22], "on the camera chip the backscattered optical signal is converted to an electronic signal and immediately correlated with the original reference signal at several phase shifts". Kolb et al. give a more intuitive description saying that the correlation signal is sampled. The sampling is realized by correlating the reflected signal with $N = 4$ phase shifted reference signals. In practice, these phase shifts are defined as

$$\alpha_n = \frac{2\pi n}{N}, \quad n = 0, \ldots, N-1.$$

The on-chip correlation leads to N raw intensity values

$$I_n = \frac{1}{t_1' - t_0'} \int_{t_0'}^{t_1'} R(t - \alpha_n)\, S(t + \varphi)\, dt$$

where φ is the phase we want to extract. The interval $T = t_1' - t_0'$ represents the integration time. This integration time is comparable to the exposure time in conventional photography. The operator of the device has to define it for every measurement which is a crucial step as explained later in this section.

We can now shift the interval in order to simplify the formula and get

$$I_n = \frac{1}{t_1 - t_0} \int_{t_0}^{t_1} R(t)\, S(t + \alpha_n + \varphi)\, dt \tag{3.1}$$

while $T = t_1' - t_0' = t_1 - t_0$. Note that the integration time T is always an integer multiple of a full period $T' = 2\pi/\omega$.

The integration time $T^\star$ set by the operator is four times the integration time T as the pixel samples only one phase shift during a period. It is technically possible to do all four samples during a single period but has some drawbacks as explained by Lange [65] in more detail. In practice the sampling is successive which increases the overall integration time $T^\star$, but reduces misalignment of the phase due to motion.

We further assume the emitted signal to be harmonic – which is not true in practice, however the calculations done by the firmware are based on this assumption. The thereby emerging error is further discussed in Section 3.2. For an analysis of the phase reconstruction from anharmonic functions, we refer to the work of Frank et al. [22] or the diploma thesis of Rapp [94].

In order to extract the phase information from the four raw intensity values I_n, we define the emitted reference signal as

$$R(t) = C_R + A_R \cos(\omega t)$$

and the reflected optical signal as

$$S(t+\alpha_n+t_d) = C_S + A_S\cos(\omega t+\alpha_n+\varphi).$$

Putting this into the correlation equation 3.1 gives four terms

$$\begin{aligned}
I_n &= \frac{1}{t_1-t_0}\int_{t_0}^{t_1}\left(C_R+A_R\cos(\omega t)\right)\left(C_S+A_S\cos(\omega t+\alpha_n+\varphi)\right)dt \\
&= \frac{1}{t_1-t_0}\int_{t_0}^{t_1} C_R C_S\, dt \\
&+ \frac{1}{t_1-t_0}\int_{t_0}^{t_1} C_S A_R\cos(\omega t)dt \\
&+ \frac{1}{t_1-t_0}\int_{t_0}^{t_1} C_R A_S\cos(\omega t+\alpha_n+\varphi)dt \\
&+ \frac{1}{t_1-t_0}\int_{t_0}^{t_1} A_R A_S\cos(\omega t)\cos(\omega t+\alpha_n+\varphi)dt.
\end{aligned}$$

The first term is substituted to

$$\frac{1}{t_1-t_0}\int_{t_0}^{t_1} C_R C_S\, dt = C_R C_S = C.$$

The second and third term both become zero as the integration time $t_1 - t_0$ is much larger than the inverse of the modulation frequency as pointed out by Rapp [94]. With a further substitution $A_R A_S = A'$ we derive

$$I_n = C + \frac{A'}{t_1-t_0}\int_{t_0}^{t_1}\cos(\omega t)\cos(\omega t+\alpha_n+\varphi)dt.$$

Now we use the Euler identity to factorize this equation again into four terms

$$\begin{aligned}
I_n &= C + \frac{A'}{t_1-t_0}\int_{t_0}^{t_1}\frac{1}{2}\left(e^{i\omega t}+e^{-i\omega t}\right)\frac{1}{2}\left(e^{i\omega t}e^{i\alpha_n}e^{i\varphi}+e^{-i\omega t}e^{-i\alpha_n}e^{-i\varphi}\right)dt \\
&= C + \frac{A'}{t_1-t_0}\int_{t_0}^{t_1}\frac{1}{4}\left(e^{i\omega t}e^{i\omega t}e^{i\alpha_n}e^{i\varphi}\right. \\
&+ e^{-i\omega t}e^{i\omega t}e^{i\alpha_n}e^{i\varphi} \\
&+ e^{i\omega t}e^{-i\omega t}e^{-i\alpha_n}e^{-i\varphi} \\
&+ \left. e^{-i\omega t}e^{-i\omega t}e^{i\alpha_n}e^{i\varphi}\right)dt.
\end{aligned}$$

From the second and third term the $e^{-i\omega t}$ parts are canceled out. Hence the whole equation can be reorganized to

$$\begin{aligned} I_n &= C + \frac{A'}{t_1 - t_0} \int_{t_0}^{t_1} \frac{1}{4} \big(e^{2i\omega t} e^{i\alpha_n} e^{i\varphi} + e^{-2i\omega t} e^{-i\alpha_n} e^{-i\varphi} \\ &+ e^{i\alpha_n} e^{i\varphi} + e^{-i\alpha_n} e^{-i\varphi} \big)\, dt \\ &= C + \frac{A'}{2(t_1 - t_0)} \int_{t_0}^{t_1} \frac{1}{2} \big(e^{2i\omega t} e^{i\alpha_n} e^{i\varphi_d} + e^{-2i\omega t} e^{-i\alpha_n} e^{-i\varphi} \big)\, dt \\ &+ \frac{A'}{2(t_1 - t_0)} \int_{t_0}^{t_1} \frac{1}{2} \big(e^{i\alpha_n} e^{i\varphi} + e^{-i\alpha_n} e^{-i\varphi} \big)\, dt. \end{aligned}$$

The Euler identity again gives two integrals

$$\begin{aligned} I_n &= C + \frac{A'}{2(t_1 - t_0)} \int_{t_0}^{t_1} \cos(2\omega t + \alpha_n + \varphi) dt \\ &+ \frac{A'}{2(t_1 - t_0)} \int_{t_0}^{t_1} \cos(\alpha_n + \varphi) dt. \end{aligned}$$

The first one becomes zero again because of the integration time being an integer multiple of the period. Removing this term and computing the remaining second integral – which is constant – leads to

$$\begin{aligned} I_n &= C + \frac{A'}{2(t_1 - t_0)} \cos(\alpha_n + \varphi)(t_1 - t_0) \\ &= C + \frac{A'}{2} \cos(\alpha_n + \varphi). \end{aligned}$$

The last step is to derive the formulas for the phase φ in order to be able to compute the distance values as intended. Therefore, we finally substitute $\frac{A'}{2} = A$ and compute the raw intensity values for each phase shift $\alpha_n = 2\pi\frac{n}{N}$ with $n = 0, 1, 2, 3$. From

$$\begin{aligned} I_0 &= C + A\cos(\varphi) \\ I_1 &= C + A\cos(\frac{\pi}{2} + \varphi) \\ I_2 &= C + A\cos(\pi + \varphi) \\ I_3 &= C + A\cos(\frac{3\pi}{2} + \varphi) \end{aligned} \tag{3.2}$$

we derive from trigonometric equalities

$$\begin{aligned} I_0 &= C + A\cos(\varphi) \\ I_1 &= C - A\sin(\varphi) \\ I_2 &= C - A\cos(\varphi) \\ I_3 &= C + A\sin(\varphi). \end{aligned} \tag{3.3}$$

From these four raw intensity values that are captured by the camera for each pixel, we can further derive formulas to compute three images: phase, intensity and amplitude.

Phase Solving this equation system for the phase φ is realized by building the differences in order to cancel the intensity C

$$I_3 - I_1 = -2A\sin(\varphi)$$
$$I_2 - I_0 = -2A\cos(\varphi)$$

and dividing them

$$\frac{I_3 - I_1}{I_2 - I_0} = \frac{-2A\sin(\varphi)}{-2A\cos(\varphi)} = \tan(\varphi).$$

This finally holds

$$\varphi = \arctan(\frac{I_3 - I_1}{I_2 - I_0}). \tag{3.4}$$

As mentioned by Rapp [94], it is necessary to take care that the results span the full unambiguity range $[0, 2\pi]$ as it is done by the arctan2 function in programming languages like MATLAB.

Intensity Besides the phase, it is possible – and realized in the firmware of most ToF cameras – to compute an intensity and an amplitude image. The intensity C is simply computed as the sum of all four raw intensity values makes

$$I_0 + I_1 + I_2 + I_3 = 4C$$

because the second terms from the equations 3.3 sum up to zero. The final intensity value

$$C = \frac{1}{4}(I_0 + I_1 + I_2 + I_3) \tag{3.5}$$

results in an image that resembles conventional camera images. However it is distorted as the reflection properties in the NIR range differ from them in the visible spectrum. In terms of image quality these intensity images are not comparable to existing camera technologies for video or movie production. Besides the missing color, the intensity images from PMD cameras have a very low spatial resolution and their image aesthetics suffer from the strong direct illumination of the scene by the NIR emitters.

Note that the PMD[vision] CamCube 3.0 measures four additional raw intensity images. These images hold the inverted signal and therefore allow to compute an image that does not contain any signal information [72]. We refer to the dissertation of Lange [65] for technical details. This enhances the image quality because the influence of the direct lighting is reduced.

Amplitude Another property of the correlated signal is the amplitude. This value is important as it is an indicator whether the sensor is saturated or not. It can be extracted from the differences

$$I_3 - I_1 = -2A\sin(\varphi)$$
$$I_2 - I_0 = -2A\cos(\varphi)$$

by squaring and adding them leading to

$$(I_3 - I_1)^2 + (I_2 - I_0)^2 = 4A^2(\sin^2(\varphi) + \cos^2(\varphi)) = 4A^2$$

as $\sin^2(\varphi) + \cos^2(\varphi) = 1$. The amplitude is finally defined as

$$A = \frac{\sqrt{(I_3 - I_1)^2 + (I_2 - I_0)^2}}{2}. \tag{3.6}$$

Summary It has been described how and which information can be extracted from a ToF sensor. The resulting three images form the basic data that is used in all depth imaging applications utilizing a ToF camera. In the following section, we discuss the reliability of this data. We reflect assumptions that result in various errors and discuss their effect on the application of the data in the area of CG. Besides describing these error sources several approaches to correct the error are presented.

3.2 Errors and noise

This section provides an overview about the existing and examined systematic errors of PMD based ToF cameras.

Preliminaries While the term *error* usually refers to the absolute difference $\Delta d = |d - d^\star|$ between a measured distance d and the true distance $d^\star$, we distinguish between spatial and temporal errors. ToF cameras do neither provide temporally nor spatially constant measurements. The temporal error is often called repeatability or repetition error. It can be indicated by capturing static scenes over a period of several (usually approx. 50) frames and calculating the standard derivation.

In order to determine the spatial error, it is necessary to define a *ground truth*. There is a general concept in order to indicate this error. It needs a specialized setup and are therefore not easy to implement. Generally, a setup is constructed where the ground truth for the distance measurement can be defined. On one hand, this can be done by using a measurement system which has smaller error and a higher depth resolution – researchers have therefore used a laser scanner [98] or laser range finders [93]. On the other hand, sensor and target can be placed at a known distance to the sensor. This method is more widespread as the setup is inexpensive and easy. Rapp [94] as well as Kahlmann et al. [51] used a high accuracy distance measurement track line. While Lindner and Kolb [70] do not mention any special equipment to measure the ground truth in their first work on calibration, they used a checkerboard pattern and additional high resolution charge-coupled device (CCD) cameras in their second approach [71]. Schiller, Beder and Koch [102] also used this setting and in a conjoint approach [76] they compare two calibration methods with and without the support of a CCD camera. Note that the term *calibration* refers to the process of comparing measurements with the ground truth [120] and finding a mathematical model that corrects the measurements as precise as possible. In the remainder of this thesis, general approaches are presented that are capable of reducing the spatial and temporal error of ToF cameras. These approaches differ from related work [69, 101] by avoiding the necessity to capture a huge amount of ground truth data. In addition, our approaches aim on extending the range of application instead of solely increasing the precision of the measurements.

In most other theses [28, 65, 69, 104] about ToF imaging, the authors depict existing errors by first describing the error, then explaining the sources and finally presenting solutions to reduce the error. In this thesis, we look at the underlying assumptions that lead to the specific error in order to facilitate a deeper understanding of their occurrence for the reader. We further discuss the assumptions in the context of possible applications and try to give practical examples whenever it is appropriate.

All of these errors occur from the fact that the distance calculation is based on certain assumptions. However, these assumptions are not valid in practice and lead to minor and mayor errors. In the following, the errors are presented in coherence with the underlying assumptions. Note that not all these assumptions are claimed by the camera developers explicitly. Some of them are just implied from how cameras work in general.

Assumption: The reference signal is sinusoidal. Unfortunately, it is technically very difficult to create a completely harmonic optical oscillator. Nevertheless the firmware of the PMD camera assumes such a completely harmonic reference signal and computes the depth values as described in the previous Section 3.1. In practice, the reference signal is not harmonic and hence it contains higher order frequencies or *anharmonics*. Anharmonics are the difference (deviation) of the modulated light signal sent out by the LED arrays from a sinusoidal signal. As shown in great detail by Rapp [94], this leads to a so called *wiggling error*. He calls this behavior wiggling, as the error changes like a wave in dependence of the object distance.

There are two approaches to reduce this spatial error. The first approach is to reduce all errors independent from their origin by calibration. Here, a lot of measurements of known scenes are recorded in order to find a function that maps measured values to true distance values. There are several variations of this function in the literature. Lindner et al. [76] give an overview about these complete approaches.

The second approach is on the reduction of only the wiggling error. It is based on the measurement of the outgoing reference signal which is assumed to be sinusoidal. Therefore a photo diode and an oscillator that work in the Terahertz (THz) range are necessary. Rapp [94] has compared the outputs of several commercially available sensors in 2007. He shows that all signals contain higher frequencies and hence considering this in the phase reconstruction computation leads to improved accuracy. While Rapp proposes to change the demodulation scheme by using a higher order Fourier series instead of the assumption of sinusoidal signals, Lindner et al. [74] claim that the wiggling error can be reduced by combining the standard sinusoidal demodulation scheme with a rectangular one. Here, the reference signal is assumed to be rectangular and hence the correlation function turns out to be triangular. This allows a reconstruction scheme that needs only a low number of reference images while being fast to compute. Unfortunately, this approach turns out to be less effective than complete approaches that rely on a calibration model that has been obtained from a larger number of reference images like the one presented earlier by Lindner and Kolb [70].

The assumption that the reference signal is harmonic is ambiguous. On one hand, the absolute distance measurements are less reliable while the computation cost of reconstructing the depth from the raw phase signal is low. This results in high frame rates at the cost of precision. In addition, the production costs would rise if the illumination units are manufactured in a different way that allows the emission of harmonic reference signals. This trade-off is directly reflected in the application. PMD cameras are used where fast response times are crucial like in interaction tasks. Interestingly, the low precision is also no issue in car safety applications like pedestrian detection. Here, it is not important whether the pedestrian is 2.5 m or 2.7 m away. The sensor has just to determine whether there is a pedestrian or not and whether the emergency brakes have to be activated or not. However, if the sensor is to be used to control the distance to a heading car, these 20 cm can make the difference.

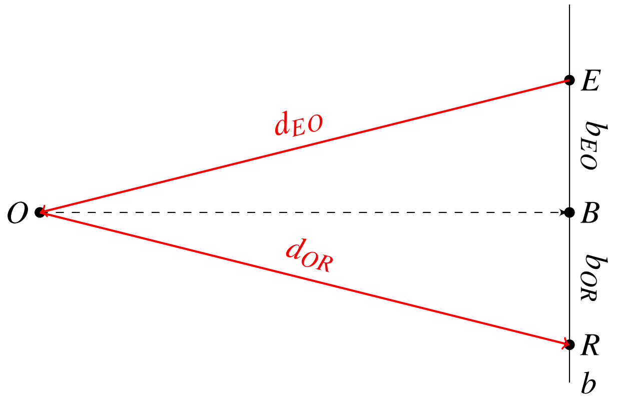


Figure 3.3: Illustration of an object point O and its projection B onto the line b defined by emitter E and receiver R.

Assumption: Emitter and receiver are at same position. The reconstruction scheme used by ToF cameras assumes that the signal emitter is located at the same position as the receiver. From this, another error due to the near field effect results. This spatial error origins in the wrong assumption that the path of the light is twice the distance between sensor and object. If receiver and emitter are not equally positioned in space, the path is along two edges of the triangle between emitter E, object O and receiver R as depicted in Figure 3.3. Hence the error depends on the configuration of this triangle. While the offset b between emitter and receiver is fixed, the object position varies for each pixel. We define the near field error

$$\delta_{nf} = |d_{EO} - d_{OR}|$$

as the difference between the distances emitter-object d_{EO} and object-receiver d_{OR}. We can figure out that this error depends on the projection of the objects position onto the base line b defined by emitter and receiver and the objects distance to this line d. We call the dropped perpendicular foot B. From trigonometry we get

$$d_{EO}^2 = d^2 + b_{EO}^2$$

and

$$d_{OR}^2 = d^2 + b_{OR}^2,$$

where b_{EO} and b_{OR} are the line segments between emitter E resp. receiver R and the dropped perpendicular foot B. See Figure 3.4 for a two-dimensional (2D) abstraction of this issue. Reconfiguring the difference of these equations leads to

$$\delta_{nf} = |\frac{b_{EO} - b_{OR}}{d_{EO} + d_{OR}}|.$$

This discloses that the error is zero for objects placed centrally in front of the sensor and increases for objects placed off the central viewing direction. Figure 3.5 illustrates this effect. Here, a plain wall is captured with a PMD[vision] CamCube 3.0 camera. The standard deviation for each pixel, which gives insight to the repeatability of the measurement that implies reliability decreases toward the border as the near field error δ_{nf} increases. Note that according to the definition, the error decreases for larger distances.

As a setup where receiver and emitter are placed at the same location is impractical and only realizable by the use of expensive beam splitters. Figure 3.1 on page 22 depicts that the LED are usually placed around (Swissranger) or beside (PMDTec) the sensor. As the receiver contains conventional optics its field of view is limited. This restriction also limits the range of possible object position relative to the sensor and hence reduces the maximal error. In practice, these sensors provide reliable data for distances larger than about 1 m.

Besides this, the PMD[vision] CamCube 3.0 (see Figure 3.1 on the right) allows to place the illumination units off the sensor. In this case, the reconstruction scheme can be reconfigured and calibrated to the new setting.

Furthermore, the assumption of emitter and receiver at the same position becomes even less valid, as both the receiver and especially the emitter are not single

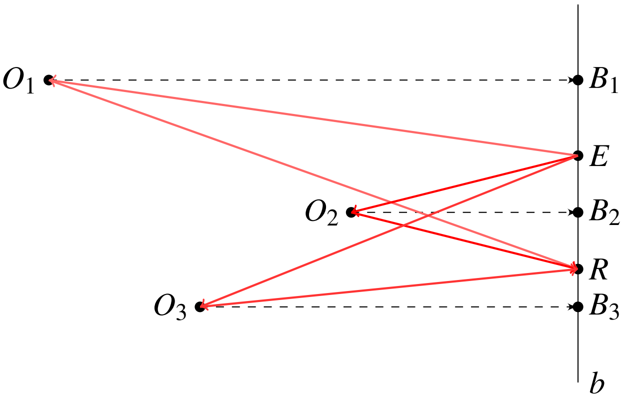


Figure 3.4: Illustration of three object points O_1, O_2, O_3 and their projections B_1, B_2, B_3 onto the line defined by emitter E and receiver R.

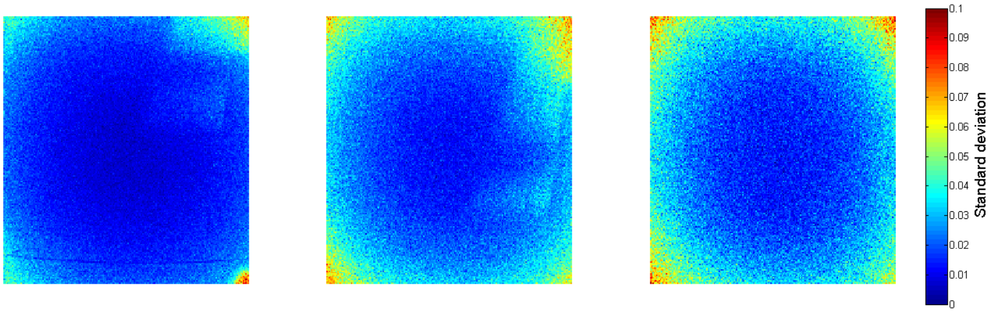


Figure 3.5: Standard deviation for captures of an office wall at 165 cm (left), 135 cm (middle) and 70 cm (right).

points. Usually LED arrays are utilized because a single LED does not provide enough power. In addition, the receiver is commonly built on a chip that has a certain area and does not equal the pinhole camera model the calculations are based on. In order to overcome these errors, PMD sensors have been simulated by Keller et al. [55] in order to develop improved sensor designs.

In the medical imaging context, Penne et al. [90] embed a ToF camera into an endoscope. Here, the illumination unit is placed externally and therefore a distance calibration is necessary. The authors describe a calibration process that includes the capture of a uniformly good reflecting object like a sheet of paper in order to correct the offset between illumination unit and sensor. In addition, some image processing steps like a bilateral filter are performed in order to reduce the error by smoothing the resulting images.

These examples show that the near-field error due to the mentioned assumption of emitter and receiver being at the same location can easily be eliminated in a simple calibration step.

Assumption: All the sensors pixel gates are equal. As the sensitivity of a sensors pixel depends on the amount of electrons and positrons in the pixel area, there are small differences between these elements. Due to charging of each pixel element there occurs a specific error which is constant in time. This constancy allows a simple calibration procedure. The error is well known from photography and occurs in any digital CCD or CMOS sensor. There are usually two main sources known as dark signal non-uniformity (DSNU) and photo response non-uniformity (PRNU). While PRNU should not have any influence, the DSNU arises from a phase delay due to different capacitance of each pixel as explained by Schmidt [104] in more detail.

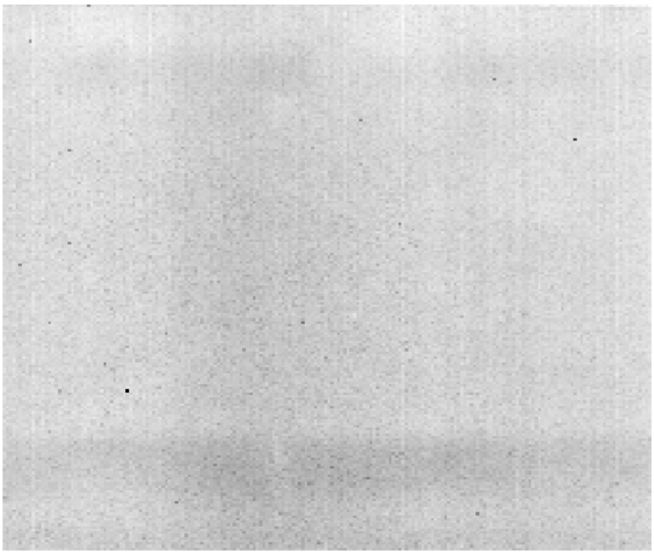


Figure 3.6: Black image captured by a PMD[vision] CamCube 3.0 camera.

However, the error can be reduced or eliminated. A simple fixed pattern noise reduction approach as described by Schmidt [104] as well as Lindner et al. [76] leads to improved results. Here, the sensor is shut by some opaque object so that no light falls onto the sensor. This should lead to a zero intensity image in theory. When recording the raw intensity values for some time and computing an time-averaged image gives a so called *black image* that can be subtracted from the raw intensity values before reconstructing the distance values. Figure 3.6 shows an exemplary black image.

Assumption: Every captured object point has equal reflection properties. The distance measuring differs from ground truth in a non-linear behavior from the amount of incident light. This amount of light depends on the reflectivity of the scene objects. Figure 3.7 shows a rendering of a captured checkerboard. In order to get rid of this error, a calibration has to be done. While Frank et al. [23] argue that the error derives from the fact that the distance measurement error is proportional to the amplitude and hence can be reduced by adaptive filtering [22], Lindner et al. [76] claim that there is no explanation available yet and provide calibration methods based on experimental data.

We will address this problem again in Chapter 6. We show that the determination of the reliability of each pixel allows to fuse several images with varying integration times. This fusion process resolves this issue.

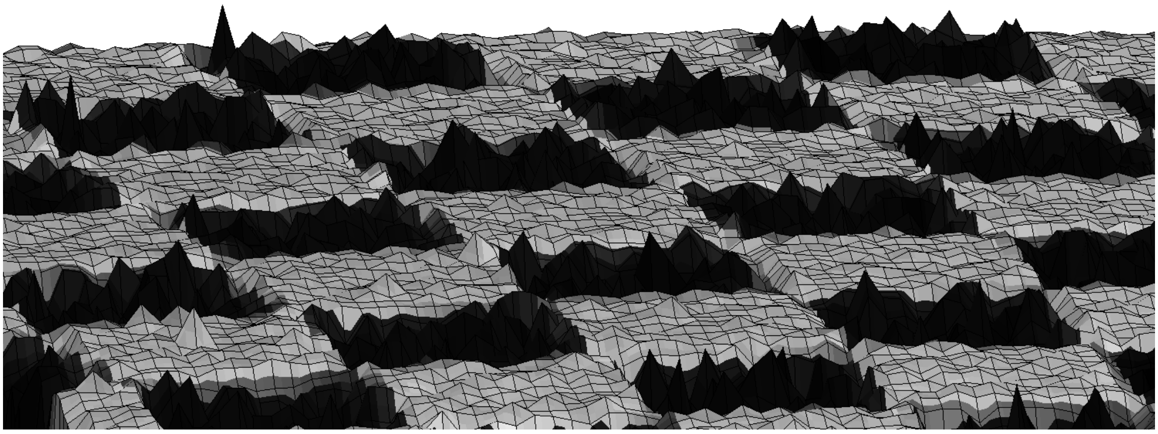


Figure 3.7: Checkerboard patterned surface.

Assumption: The distance is constant inside one pixel. Each pixel does not measure the distance to a single point but gathers the incoming light reflected from a small area which is called the solid angle. This area is assumed to be at the same distance from the object. This is not the case in practice. If the pixel covers an area including a depth discontinuity, the collected photons have traveled diverging distances and hence their time-of-flight is different. This means that the raw intensity values measured on the chip are sampled from several (usually just two) phase shifted signals.

In practice, this usually results in so called *flying pixels*. Figure 3.8 displays this effect on an exemplary office scene. The flying pixels occur between the monitor and the wall and are highlighted with a surrounding polygon. Keller and Kolb [54] have simulated this behavior by super-sampling. Their results resemble the real world data so that the existence of flying pixels can be fully explained by the violation of the stated assumption.

Note that this problem is specific for range imaging methods that utilize digital image sensors that do spatial sampling and quantize the information. The key problem for ToF imaging is that this error is hard to detect. Huhle et al. [45] propose the usage of additional color information from an external camera to detect

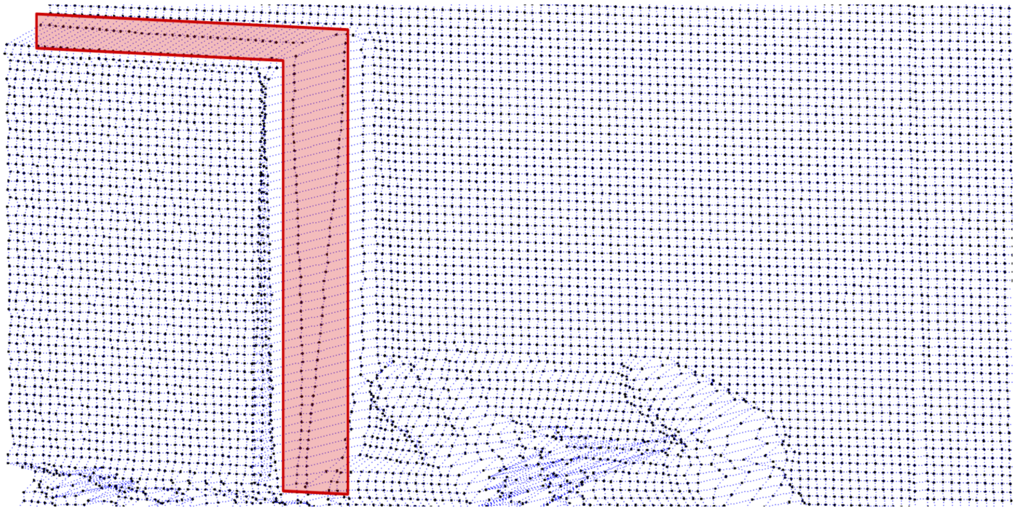


Figure 3.8: 3D point cloud of a monitor in front of a wall with flying pixels inside the polygon.

those pixels, while Lindner et al. [73] use a gradient-based resampling approach to preserve sharp edges. Swadzba et al. [109] propose a method that removes all pixels that do not have a sufficient number of neighbors at a similar distance. This approach has been shown to be capable of enhancing the results of a iterative closest point (ICP) algorithm in order to register point clouds. However, Reynolds et al. [98] have shown recently, that the elimination of flying pixels can be further improved. They assign each pixel with a confidence value that results from a supervised learning approach. They capture ground truth images from various scenes using a laser scanner and train a random forest that holds confidence values for each feature. As a feature, they define a vector containing local, spatial and global features extracted from the distance as well as the amplitude and intensity images of the ToF camera. Their experiments show that the distance itself and the Laplacian of distances are the most important features.

In the remainder of this thesis, we will depict how both the combination with stereo imaging and the fusion of several exposures reduce the issues with flying pixels.

Assumption: There is no motion for the time of integration. A similar error occurs if the captured objects or the camera itself are moving. Then, the probability that the distances – and the reflectivity – inside the solid angle of a pixel are constant is strongly decreased. The error occurs mainly around depth discontinuities because here the difference inside the integration time window varies the most.

Note that this effect is well known from photography. Here, it is called motion blur. Motion blur makes detection and recognition applications challenging as the blurred images contain less sharp features and hence less information than focused images. In the field of computational photography, there are several approaches to overcome this problem. Most prominently the fluttered shutter approach by Raskar et al. [95]. Here, instead of a permanently opened shutter during exposure, the shutter is opened and closed in a specific pattern that allows the reconstruction of an less blurred image while keeping the amount of incoming light high enough.

Further motion blur occurs, when the object moves during the capture of the four raw intensity values from which the depth and amplitude values are estimated. A wrong raw value leads to completely wrong depth estimations. According to Lindner and Kolb [72], the motion artefacts can be reduced by registering the raw intensity images with optical flow methods before extracting the distance information. This approach is possible as the raw intensity samples are not taken during the same period but successively. This approach has been applied in the setting of the surveillance of a conveyor belt by Hussmann et al. [46]. Here, knowing the motion direction allows the compensation of artefacts and hence a reliable extraction of distance values.

Assumption: There is no other NIR source but the emitter. This assumption is obviously not valid in the presence of sunlight. The additional light is also received by the sensor and may lead to a saturation. As primary methods to increase the signal to noise ratio (SNR) between received signal and sunlight, optical bandpass filters are utilized and the LED are run in a burst mode. Additionally, it is possible to detect whether the received light is part of the emitted signal or not. Therefore the sensors chip needs two readout cycles that correlate with the expected incoming signal. As the noise signal is uncorrelated both readout sides receive an equal amount of noise information that can be rejected. Basically, this is the same principle as differential signaling [121]. This is implemented by the suppression of background illumination (SBI) technology in PMDTec cameras as described by Möller et al. [83].

Assumption: The light signal is reflected directly. As most real world objects reflect light in different directions, many photons emitted by the illumination unit find an indirect path back to the sensor. As the path length differs from the direct way, the phase measurement is wrong in this case. However, the ratio between direct and indirect lighting is so high that the effects are negligible.

Recently, there is some research about how to exploit these reflections by the camera culture group at the MIT media lab lead by Raskar. They use the reflections both to allow a camera to look around the corner [58] and to reconstruct reflectance properties of objects from a single photograph [84]. This development is in an very early state and suffers from the already mentioned ratio of direct and indirect light.

Assumption: The sensor works under all conditions. There is a minor issue, that PMD sensors are influenced by the internal and external temperature.

Kahlmann et al. [51] present the results of some measurements with a Swissranger SR-2 camera at different external temperatures. Theses results show a drastic differences of roughly 30 cm between −10 °C and 30 °C. This issue becomes crucial if the camera is utilized in non-standard environments. If the sensor is embedded into a car close to the engine, temperatures around 80 °C are prevalent. Although the thermal noise is uncorrelated, this problem cannot be solved by the SBI technology as described above as the temperature affects the illumination unit and hence changes the reference signal. This leads to the problems that have been discussed above in the context of anharmonics.

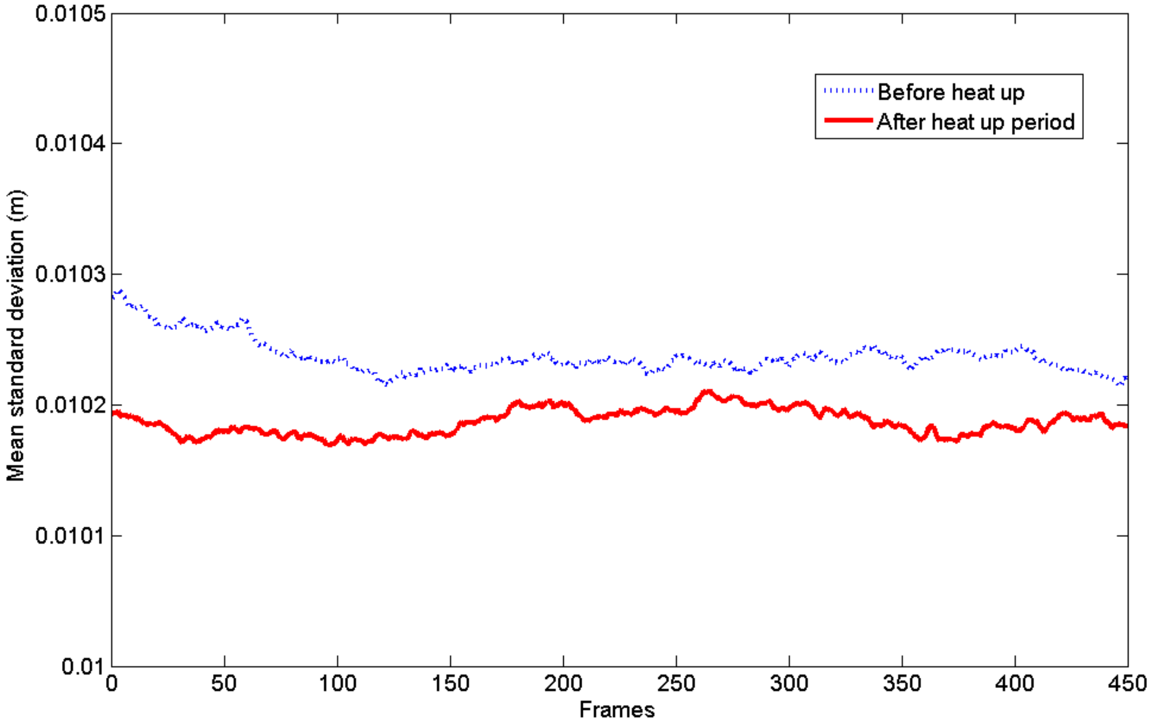


Figure 3.9: Comparison of mean standard deviation of depth values for a static scene.

Further experiments by Kahlmann et al. [51] show that the camera has to be heated up before exact measurements are possible. In order to illustrate that this effect is strongly reduced but still remarkable, 500 frames of a static scene have been captured with a PMD[vision] CamCube 3.0 camera. Once, directly after switching on the camera and a second time after a heating up period of roughly 20 minutes. In order to depict the temporal error, we compare the standard deviation of the depth values of 50 successive frames for each pixel. Figure 3.9 displays the mean value over all pixels for both measurements. It shows that the standard deviation is slightly reduced but only in a range of some millimeters.

Assumption: The longer the measurement, the higher the precision. One last issue comes with the effect that the measured distances are not independent

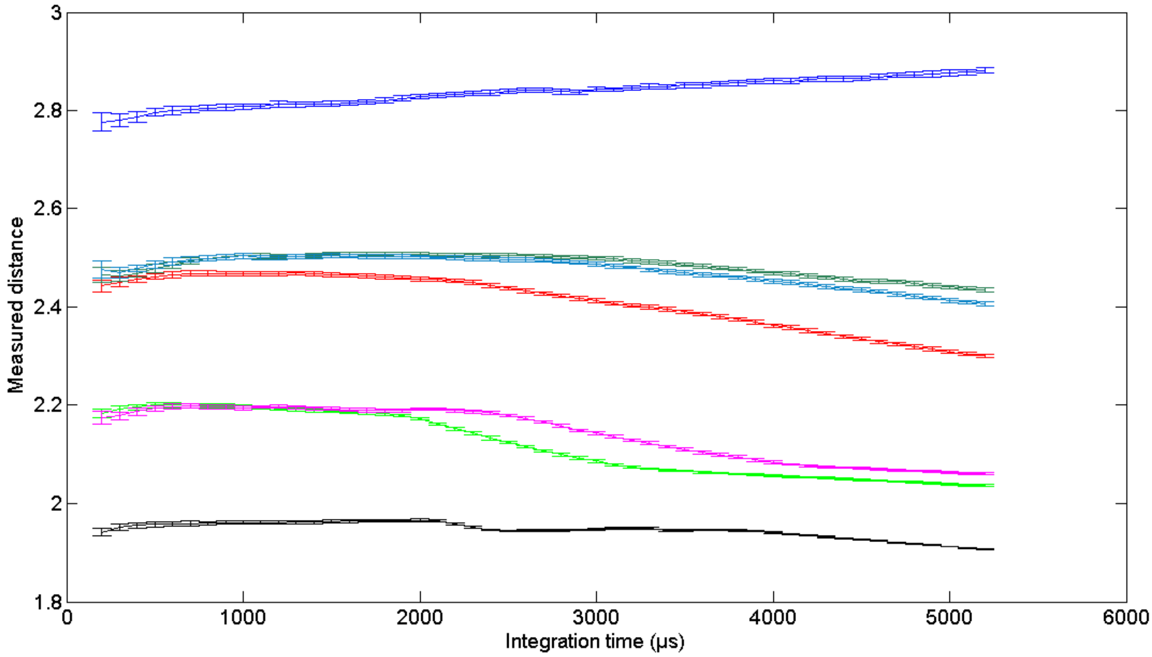


Figure 3.10: Measured mean distance of the center pixel against the integration time.

from the integration time. While for older camera models Kahlmann et al. [51] have shown large differences in the measured distances, this effect is reduced in recent camera models like the PMD[vision] CamCube 3.0. However, the measured distance is not constant for different integration times. We measured the distance of the center pixel for a static scene with varying integration time in the range of 100μs to 6 000μs. We plot the mean distance over 50 frames against the integration time in Figure 3.10 for several static scenes with different distances of the center pixel. The ideal measurement should lead to a horizontal line in the plot. We included error bars for the standard deviation of the mean distances. They indicate that the variation decreases with increasing integration time.

While the source of this error remains unclear, we provide a solution to overcome the problem of setting the right integration time. It is based on an idea from computational photography by Mertens et al. [80] and described in detail in Chapter 6.

3.3 Conclusion

In the last section, numerous error sources and other issues with ToF cameras have been described. Several assumptions have been stated that do not hold in practice. While some of these assumptions do not lead to crucial errors in the measurements, others lead to significant issues that avoid the application of ToF cameras in certain fields.

This makes it indispensable to enhance the data before processing it further or apply in existing frameworks. This outcome forms the foundation of the key contributions of this thesis. We argue that it is necessary to deal with the drawbacks of ToF cameras before further processing the data.

One additional drawback that until now has been neglected is the missing color information from the sensor. There is no high resolution video camera embedded that allows to capture color and light from the scene as a conventional photo camera does. It is straightforward to attach a high resolution photo camera beside the ToF camera. The imaging nature of the ToF camera suggest itself to register the imagery of both cameras with standard methods as it is done for two conventional cameras in practically all stereo vision setups. We further argue that attaching two instead of only one photo camera does not increase the burden, however it allows to capture depth information from stereopsis as well. As described in Chapter 2, stereo depth imaging has very different properties than ToF imaging. In the next chapter, we describe a proof-of-concept that combining a ToF camera with a stereo camera setup is implementable and leads to improved depth imaging results. Note that the chapter has been previously published in large extent as [36] and [37].

Chapter 4

Combining Time-of-Flight and stereo images

4.1 Introduction

Reliable and fast depth imaging for real world scenes is a difficult problem. A variety of methods to solve this task has been presented in Chapter 2, however, each of them has its strengths and weaknesses. In particular, passive methods rely on feature matching, which is time-consuming and might fail if no features are present. Active techniques overcome this problem, yet at the price of being sensitive to the reflection properties of the elements in the scene and typically higher cost of the devices.

As pointed out in the Chapter 3, a new type of ToF sensor has extended this spectrum: PMD cameras can be used to measure the phase of a modulated light source relative to its reflection in the scene. The phase directly relates to the distance. These devices are low-cost and very fast, but their spatial resolution is low and they suffer from noise, especially around depth discontinuities as explained in detail in Section 3.2.

It is interesting to compare this new approach to the dominating low cost depth approach, namely stereo. For planar patches, Beder et al. [5] find that the PMD approach is more accurate than stereo. On the other hand, details and discontinuities in intensity and/or depth decrease the performance of PMD depth measurement, while they typically increase the performance of stereo. Also, traditional cameras have a much higher resolution and stereo setups have a larger working range.

A natural conclusion is to combine the PMD approach with standard stereo photography. Few approaches consider the fusion of PMD generated depth (and possibly intensity) images with standard photography. Reulke [97] combines a single high resolution intensity image with the PMD depth image. Intensity information is exploited to steer the re-sampling of depth data into a higher resolution depth image. Kuhnert and Stommel [64] propose combining the PMD approach with an additional stereo camera. This is identical to our setup. They take the depth image of the PMD camera and a depth reconstruction from the stereo pair as independent measures of the depth in the scene. For pixel with low confidence in the computed disparities (i.e. no significant matching for the windows) they fill in the data gathered with the PMD camera. We are using a similar approach of combining PMD and stereo depth. The stereo pair is mounted symmetrically to the PMD camera, with the three centers of projection (roughly) co-linear and parallel image planes (see Section 4.2 for details). This yields three images, one depth image and two (color) intensity images. The three images are calibrated based on their intensity images using standard approaches [41, 128, 129]. Because of the low resolution and the fact that the PMD sensor rather measures intensity differences than absolute values, this calibration turns out to be not very accurate. However, even the data based on the inaccurate calibration can be fused.

We are using similarity of intensity values (over windows) from the stereo pair together with the phase information from the PMD camera, which is new compared with prior approaches. While this idea could be used in any stereo setting, we are here using a global approach based on graph cuts [62, 89, 44, 112, 126] for the reconstruction of a depth image. Note that graph cuts have recently been used for image segmentation with ToF cameras [2, 24] but typically, these approaches are computationally demanding, yet in our setting we can exploit the PMD generated depth image for restricting the domain of the volumetric grid. In Section 4.3, we explain how we use graph cut in our setting. The results of this procedure allow increasing the resolution of the PMD depth data, while keeping sharp depth discontinuities based on intensity discontinuities. An important feature of the setup is that we could record a video sequence, and then recompute the depth on a per frame basis. This yields a system with dramatically improved depth reconstruction for dynamic scenes.

4.2 Setup

We run our experiments using a PMD[vision] 19k camera with a resolution of 160×120 pixels. It is centered between two standard photo cameras. As an experiment, we have used consumer grade cameras of type Olympus SP-500 UZ. These cameras could be interchanged by standard firewire cameras in order to capture a video sequence for reconstructing dynamic scenes. All three cameras are mounted on an aluminum bar (see Figure 4.1 for an image of the cameras). The stereo pair has a base line of roughly 50 cm. The cases are mounted so that their image planes are parallel.

For calibration of this setup, we need to pay special attention to the specifics

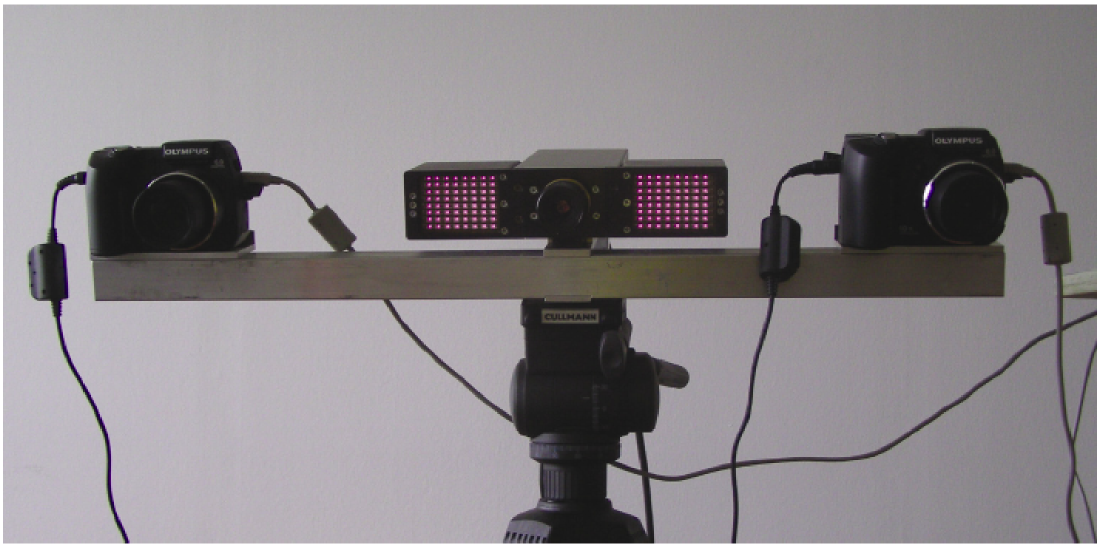


Figure 4.1: All three cameras mounted on an aluminum bar

of the PMD camera. It turns out that intrinsic calibration is not very accurate (the reprojection error is larger than a pixel) and extrinsic calibration in the classical sense would not be sufficient: extrinsic calibration usually tries to align the optical systems of the cameras, but it is unclear if zero phase shift in the PMD sensor would exactly match the optical center.

Intrinsic calibration Due to manufacturing mechanics, the intrinsic parameters of a consumer camera given by the producer are very inaccurate. Therefore, calibration is necessary. We use the algorithm of Zhang [128, 129] for the standard cameras. The intrinsic calibration of the PMD camera is more difficult, mostly because of its sensitivity to reflections of the self emitted IR light and low resolution. Another problem is that the PMD[vision] 19k camera only measures differences in capacities. Hence, there is no true intensity image available. The intensity image can be simulated by weighting the amplitudes according to their squared distances. In order to recognize feature points (e.g. corners on a checkerboard) automatically, the calibration sheet has to be placed quite directly in front of the camera. Apart from these details, we follow the ideas of Reulke [97] as well as Lindner and Kolb [70].

Note that the intrinsic calibration of the PMD camera already yields Euclidean coordinates for each pixel: the coordinates of the pixel identify a ray by means of the intrinsic transformation. The pixel contains a distance value, which identifies a point on the ray. Thus, the intrinsic transformation allows computing a set of points in $\mathbb{R}^3$ from the depth image.

Extrinsic calibration Our main idea is to use well-known techniques to perform an extrinsic calibration for all three cameras. We decided to use the OpenCV calibration methods as well as the Camera Calibration Toolbox for MATLAB that are based upon the same composition of algorithms [41, 128] and are written by Bouguet. We have found, however, that the low resolution and relatively high noise level lead to very inaccurate and alternating calibration results. In addition, there are two further issues: first, the plane of values with zero phase shift, which defines the plane with zero depth, is not necessarily coinciding with the image plane of the camera. Second, the depth measurement is affected by systematic errors as described in Section 3.2. We model this effect based on experiments by moving the camera plane along the depth axis so that depth values obtained with the camera coincide with stereo depth values.

Imaging For taking the intensity/color images, we use the application programmers interface provided by Olympus Corporation[1]. This allows setting similar parameters for both cameras. In order to get a representative depth image, we take on the order of 20 images with the PMD camera. The observation that the data values are unstable reflects the errors described in Section 3.2. Working with the mean or the median image from several images not only significantly reduces noise in the depth image of the PMD camera, but also it provides useful information on the confidence in the depth values: we use the variation of the depth values around the median value as a measure of the confidence. In particular, we use the variance of each depth value as a weight for taking into account the depth values of the PMD camera as well as for defining the domain for the graph cut algorithm. This is explained in detail in Section 4.3.

4.3 Algorithm

In the following, we describe the details of the depth computation. We follow the graph cut reconstruction by Paris et al. [89]. Our goal is to reconstruct the surface S in the object space (x,y,z) defined parametrically by the function f : $S(u,v,f(u,v))$, which can be interpreted as a depth function $z = f(x,y)$. The surface S minimizes the functional

$$\iint \left(c(S) + \left(\alpha_u(u,v)\frac{\partial S}{\partial u} + \alpha_v(u,v))\frac{\partial S}{\partial v} \right) dudv \right) \tag{4.1}$$

containing a consistency term c and two smoothing terms α_u and α_v. In the classic stereo graph cut, c is defined by stereo correspondences and, α_u and α_v by discontinuities in x and y direction in both images. The data obtained with the PMD camera allows us to refine the definition of c and α – we will describe our enhanced definition in the remainder of this section.

[1] Olympus Camedia SDK 3.4 (2004)

Defining the volume The graph cut approach works on a discrete volume. The number of voxels in x and y direction determine the resolution of the resulting depth map. The number of voxels in z direction defines the quantization of depth values. In our experiments, we have used grids of dimensions $400 \times 300 \times 100$ in width, height and depth. Note that for scenes with higher dynamics in depth the number of voxels in depth should be extended.

For all of these voxels, we build a mapping function $M : \mathbb{N}^3 \to \mathbb{R}^3$ that maps a voxel id to its 3D position in the PMD camera coordinate system (x,y,z). For this mapping, we use the intrinsic values of the PMD camera to find the x and y positions of the voxel (see previous Section 4.2). All depth steps, from the minimal depth value to the maximum depth of the PMD camera image, are filled with corresponding z values. We receive the minimal and maximal depth in the scene from the set of median depth values. In order to reduce the computational costs, a domain of interest (DOI) is defined inside this volume. The DOI is determined as follows: First, we find the voxel which is closest to the PMD depth values. Hence, the DOI contains one voxel in each depth column. If this depth value has a large variance, voxels before and after (in z-direction) are added to the domain, as well. Then, these voxels are connected in x and y direction. Therefore, we add voxels from the neighboring depth columns until we reach a connected set of voxels. Thus, the DOI can be interpreted as the voxelization of the variance controlled blurred PMD image. This is a significant reduction of computation cost for stereo graph cut algorithms, which usually have to start from the overlap of the viewing frustums of the left and the right camera.

Constructing a graph After defining the volume and the DOI, a graph is constructed according to Paris et al. [89]. This graph is connected to a source node which is placed in front of our object space (minimal depth) and a sink node behind the object at maximal depth. Basically, graph cut algorithms aim at setting up the capacities in such a way that the minimal cut describes the demanded surface. The graph contains two types of edges: consistency edges and smoothing edges. The consistency edges are inside one voxel and their capacity is determined by the measure of probability that this voxel belongs to the surface. We compute this measure as a linear combination between commonly used stereo consistency terms, here denoted as c_{stereo}, and a consistency value c_{PMD} computed from the difference of the depths of the voxel and the corresponding PMD depth. The stereo consistency is either derived from the normalized cross correlation or sum of squared distances between the corresponding regions in the left and right image. The depth difference d can then be mapped to a consistency value c_{PMD} using a convex or concave function. However, the convex term $c_{PMD} = 1/d^2 + 1$ and the concave term $c_{PMD} = max[0, 1 - d^2]$ were indistinguishable in our experiments.

Computing the smoothing edges capacities is quite similar. The stereo smoothing value α_{stereo} depends upon color discontinuities in both stereo images. For the PMD smoothing term α_{PMD}, we use depth discontinuities in the PMD median image. The terms are combined linearly, as well.

Based on the topology of the graph and edge weights, the minimal cut computed with standard graph cut algorithms determines the surface. As with other depth reconstruction approaches based on graph cut, this is the dominating factor in the computation time, and in our C++ implementation it requires a few minutes at typical input.

4.4 Results

Figure 4.2 shows our stereo input images. We chose this arrangement of cardboard boxes in order to clearly see the gradient of depth in our results. Supplementary, the untextured cardboard is a typical case where classic stereo algorithm may fail, because there are problems in identifying disparities on large uniform colored areas in the stereo image pairs.

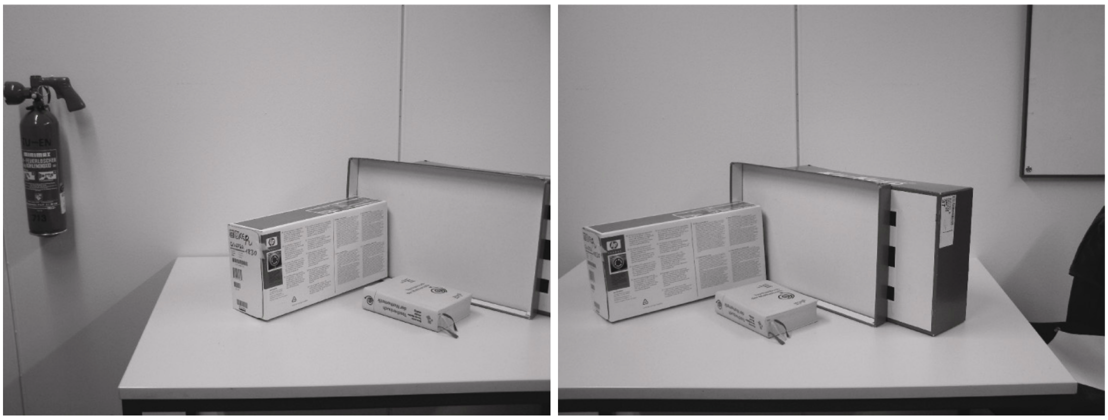
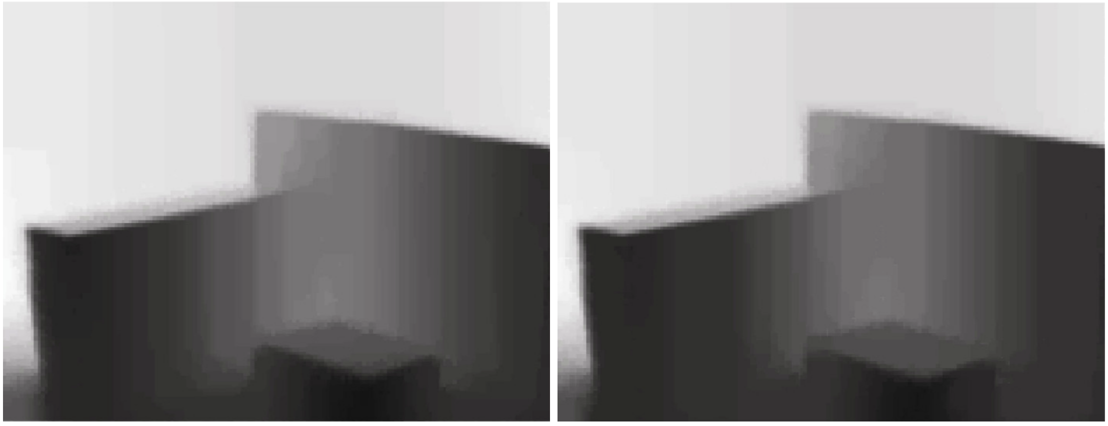


Figure 4.2: Stereo pair of images.

As mentioned in Section 4.3, the reconstruction of the surface is performed by finding a minimal cut for the constructed graph. This cut minimizes the functional given in Equation 4.1. As expected, computation times are greatly reduced by using a tighter DOI. In addition, if we extend the DOI by a dilation, the results in Figure 4.3 show that using this larger DOI does not enhance the result. However, it takes much more time to compute.

In addition to the computational savings, we were interested in whether the accuracy of the reconstructed surface also improves on using PMD depth data or stereo reconstruction alone. The characteristics of the PMD depth data would typically lead to either noisy or overly smoothed results, especially when re-sampled to higher resolution: note that each depth element is measured independently of its neighbors, resulting in noise with salt and pepper distribution. Removing this noise results in smoothing. In addition, increasing the resolution of the image also results in smoothed edges. The stereo information and the smoothing term used

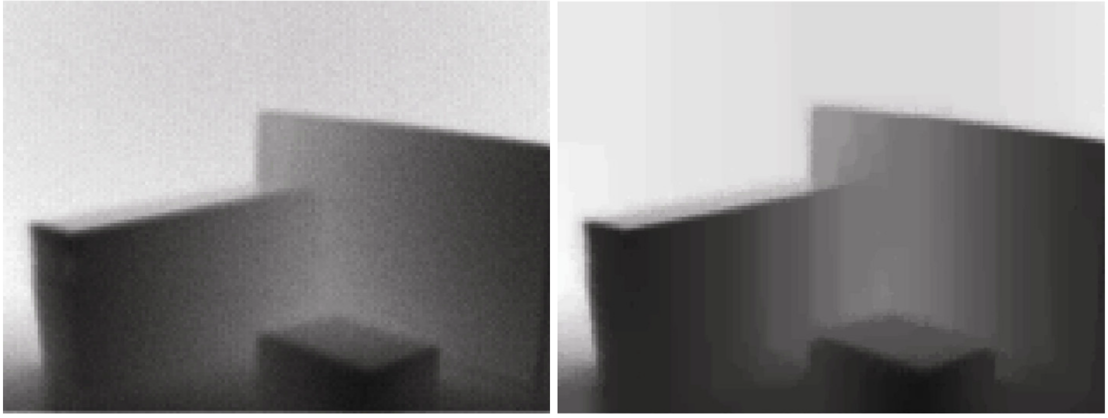


(a) regular DOI

(b) dilated DOI

Figure 4.3: Increasing the DOI leads to blurred results and longer computation time.

in the graph cut algorithm greatly reduce this noise (see Figure 4.4). In all flat surface areas inside the scene, the grainy artefacts are removed. Depth estimation is enhanced as well: in contrast to the PMD median image, the cardboard boxes are reconstructed without distortion. Their depth increases smoothly along the surface and stays constant on the vertical axis. Also, depth discontinuities are more accurately modeled by exploiting the color discontinuities in the higher resolution intensity images. This improvement can be recognized clearly. In the PMD image at the borders of the book lying in front of the cardboard boxes, we can see a slightly brighter area at the depth discontinuities – generated from inaccurate depth values, while in our resulting graph cut image, these effects are removed.



(a) PMD median filtered

(b) Graph cut smoothing

Figure 4.4: Salt and pepper noise is reduced and the depth information is enhanced.

4.5 Discussion

We have shown that the fusion of PMD data with stereo images enhances depth reconstruction in low-cost sensor systems. Our experiments are based on only roughly calibrated systems, and therefore, we do not evaluate the results quantitatively. Nevertheless, we have found the results to be better than with either system alone.

This result motivates the next step: setting up a more accurate system. This is more complicated than for usual camera based systems, as the sensing technology in the PMD camera appears to be not accurately modeled with a perspective transformation alone. Furthermore, the calibration of consumer grade zoom cameras has to be explored carefully, as well. An accurate calibration would allow us to exploit the two types of information for each voxel in a more systematic way. In addition, an accurate calibration will be necessary, if we want to start an exact evaluation about the accuracy of our system. We therefore, rebuild the system by replacing the zoom cameras with professional video cameras and known optics.

In the next chapter which has been previously published as [38], we depict how using the video cameras for the stereo pair results in a system capable of recording depth data at high frame rates and for very low cost. The characteristics of such a system would make it very attractive for a variety of applications, perhaps most prominently 3DTV as aforementioned in the introduction of this thesis.

Chapter 5

Real-time depth imaging by combining Time-of-Flight and on-demand stereo

5.1 Introduction

Real-time depth imaging is a building block in many interactive vision systems and, in particular, is necessary for enabling realistic occlusions in Augmented Reality (AR). Despite the improved speed of general purposes computing as well as development of new types of sensors, providing depth images in *real-time* continues to be a challenging problem. For the purposes of enhancing AR with convincing occlusions current approaches are limited to either reducing quality in the depth maps [103] or realizing occlusions by compositing [114]. We are demonstrating a first step toward an affordable and lightweight solution by fusing information from a low cost but also low resolution ToF range sensor with standard correlation-based stereo.

As pointed out in Chapter 2, we can distinguish active and passive approaches to real-time depth imaging. Active optical techniques involve relighting the scene and usually require an expensive and heavy setup. In Chapter 3, we depict that sensors based on theToF principle for sensing depth have become affordable and fast with the introduction of PMD.It allows depth imaging at interactive rates, but suffers from comparably low spatial resolution of the sensor and noise in the depth values, especially for surfaces with low reflectance.

Passive techniques, at least when several frames per second are required, are based on multiple views of scene captured with two or more cameras. We have decided to use a single binocular stereo camera in order to keep the setup simple and reduce processing costs. From the large number of stereo vision approaches [100] only local correlation based methods are fast enough for real-time application [17, 42]. This limits the quality of the resulting depth maps, most obviously in large featureless areas but also at depth discontinuities, where the correlation window might compare different objects because of occlusion. In a more recent work, Hirschmüller and Scharstein [43] show that stereo matching algorithms that deal with the elimination of radiometric distortion are not capable of being used in real-time applications with reasonable high video resolution.

We combine the camera systems (ToF and stereo) and fuse the data so that limitations of each of the individual sensors are compensated. Our goal is enhancing the high resolution color image from one of the stereo camera oculars with depth information, gathered from the PMD camera as well as from disparity estimation. The mapping of the PMD depth image into a color image acquired with another camera (resp. texturing the depth data with the color image) has been analyzed by Reulke [97] as well as Lindner et al. [71, 73]. Our setup combines the PMD camera with a binocular stereo camera, similar to [5, 34, 86, 64, 130, 131].

For better explaining our choice of algorithm, we need to briefly touch on the setup, calibration, and properties of the cameras (Section 5.2). The physical properties of the PMD camera give rise to the preprocessing of its data, most importantly the estimation of confidence values for each depth value (Section 5.3). Kuhnert and Stommel [64], as well as Netramai et al. [86] use a similar confidence map to choose either the depth value acquired with the PMD camera or depth from stereo – we exploit this depth/confidence map for initializing and steering a local correlation based stereo algorithm (Section 5.4), in particular by choosing adaptive windows for the correlation based on the information in both the color images and the range image.

The following sections contain illustrations and results from the diploma thesis of Mischler [82] that is based on ideas of the author of this thesis.

As described in the previous Chapter 4, we combined the ToF data from the PMD camera with high resolution images from two photo cameras, using graph cuts to find a globally optimal solution for a depth map of a single perspective. The use of graph cuts leads to computation times that are insufficient for real-time video processing. Similarly, Guðmundsson et al. [34], Zhu et al. [130], and Beder et al. [5] generate depth images by fusing ToF and stereo data. Their approaches appear to be much faster than using graph cuts, however, they target single images and provide no information on the computation times. The choices of stereo algorithm, however, indicate that they are not amenable to real-time processing in their current form. In a more recent work that is based on [130], Zhu et al. [131] state that their fusion process takes about 20 seconds.

In contrast to the approach of Koch et al. [60], we explicitly start from the restrictive setting of real-time applications, which severely restricts the choice of stereo algorithm, mostly to local correlation with fixed windows. We use the ToF information particularly to adapt the windows, as fixed windows fail at depth

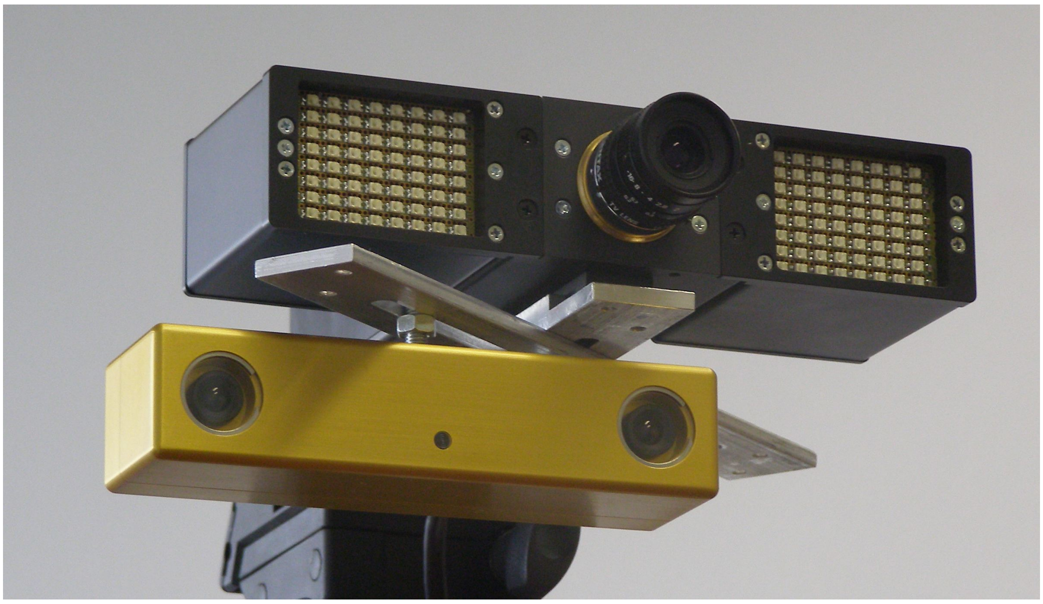


Figure 5.1: Both camera systems mounted on an aluminum bar.

discontinuities. We believe our approach yields depth images at interactive rates

- that are are more reliable than the information from the PMD camera without compromising the interactive frame rate and
- that are more accurate around depth discontinuities than real-time stereo vision approaches based on fixed window correlation.

We demonstrate our use of the system in an AR scenario for computing accurate occlusions between virtual and real objects.

5.2 Setup

We mount a compact PMD[vision] 19k ToF camera, capturing a depth range of about 7.5 m at a resolution of 160 × 120 pixels, together with PointGrey Bumblebee2 stereo camera capturing color video at a resolution of 640 × 480, on an aluminum rack (see Figure 5.1 for an image of the cameras). Both cameras are aligned to parallel viewing directions.

Calibration Sensor fusion requires registration and accurate calibration. For both the intrinsic and extrinsic calibration we use the calibration algorithm of Zhang [128, 129]. The PMD camera, however, yields intensity images that are too noisy for direct application and they are preprocessed following the ideas of Reulke [97] as well as Lindner and Kolb [70]. A relevant practical problem for the extrinsic calibration is the misalignment of optical center and zero depth plane of the PMD camera. Interestingly, Guðmundsson et al. [34] perform a stereo calibration between all pairs of cameras. We have found this to be cumbersome, because of the combination of noisy amplitude images, the mismatch between optical center and depth image for the PMD camera as well as the systematic errors in the depth measurement. We rather consider only the extrinsic calibration between the depth image from PMD camera and the systems of the color cameras.

The stereo camera color images are rectified to reduce the correspondence problem to a single line. Our calibration is accurate enough so that the depth difference between stereo system and the PMD camera is within the accuracy of the PMD camera.

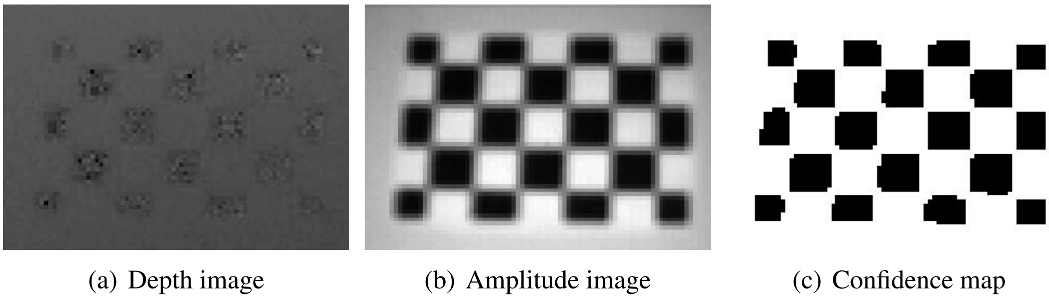
5.3 Preprocessing and confidence estimation

As explained in the last section and in Section 3.2, the quality of the depth values captured by the PMD camera depends strongly on the surface of the objects in the scene. Dark and glossy surfaces lead to artifacts as the modulated IR light is not reflected as expected. Especially when using the depth values for determining occlusions in AR applications, these artifacts become clearly visible. We process the data prior to using it with the stereo system, trying to improve the data by simply filtering and assigning confidence values to each depth value. Very low confidence depth values are replaced by interpolated values with higher confidence.

Filtering Reducing the noise or removing outliers is one obvious part of the preprocessing. Isolated outliers can be removed at minimal cost using median filtering. Through experimentation we have found that a small kernel of 3×3 pixel is sufficient. This appears to be due to the outliers being mostly isolated pixels. Larger kernels would lead to longer processing times without showing a significant improvement.

Confidence estimation As explained in the last section, in our setup, the dominant cause for systematically wrong depth estimation are objects that have bad reflection properties due to their material and color. However, this information is available in form of an amplitude image of the scene.

The first step in turning the amplitude image into a confidence map is applying a 3×3 median filter, similar to the process for the depth image. Although the resulting image is not truly bimodal, applying a simple binary threshold yields a binary confidence map classifying depth values as either valid or invalid. Note that such a definition of a confidence value from the amplitude has been applied in many other approaches [61]. In addition, Frank et al. [22] show that the amplitude value is an optimal indicator for the confidence of the range data. Figure 5.2 highlights



(a) Depth image (b) Amplitude image (c) Confidence map

Figure 5.2: Checkerboards are difficult to reconstruct using PMD range sensing, because of the insufficient amount of light reflected by the black areas. The acquired depth image (a) clearly holds wrong depth values in these black areas. The amplitude image (b) can be used to compute a confidence map (c), which is thresholded to classify depth values as valid (white) or invalid (black).

the problems resulting from low reflectivity at the example of a checkerboard and shows the resulting classification of valid and invalid depth values (while the 3D reconstruction from the wrong depth values can be seen in Fig. 5.4(a)).

Interpolation In all regions marked as confident by the binary map, we will use depth values for initializing disparities. In most cases the depth values in insufficiently reflecting areas are very distant from the estimated ground truth and it is better to assume continuity in depth.

As in our case the interpolated depth values will later be corrected using the stereo information, we opt for a simple approach that is as fast as possible: the depth image is scanned across horizontal lines. When an invalid segment is encountered it is replaced by either a line connecting the two valid depth values at the boundary of the segment or a line with constant depth values at the boundaries of the image (see Figure 5.3). For textured planar surfaces (such as the

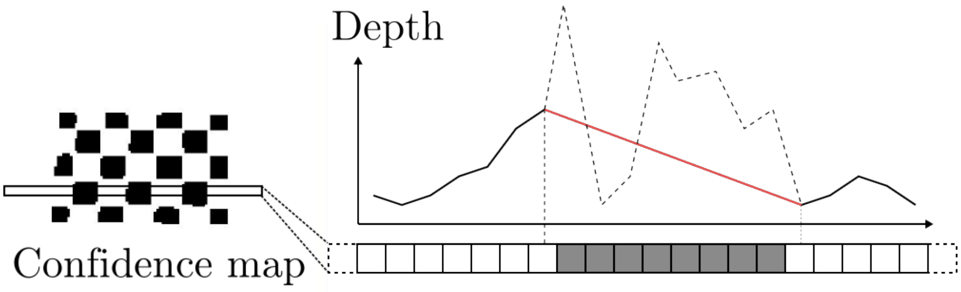
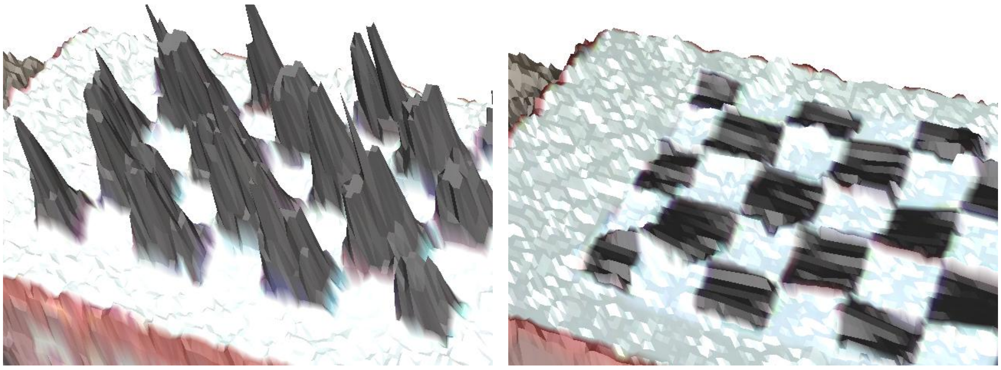


Figure 5.3: Scanline interpolation of the PMD data using the confidence map. The red line is the interpolated depth, the dashed line is the original unreliable PMD depth.



(a) Original image

(b) With interpolation

Figure 5.4: 3D reconstruction of the scene from the depth image (and using the intrinsic camera geometry). The left image shows the reconstruction from the data in Figure 5.2(a) and the right image uses linearly interpolated depth values for elements with low confidence.

checkerboard, see Figure 5.4(b)) this provides a reasonable estimate; if objects of low reflectance differ in depth from the surrounding they will be assigned wrong depth values, which will be corrected in the stereo part of the algorithm.

5.4 Algorithm

In our exemplary AR applications with dynamic occlusions we need to enhance one color image from the stereo camera with depth information. We compute depth values at the resolution of the PMD camera. The algorithm we suggest is equally applicable for computing depth at higher resolution or textured surfaces in other views.

The main steps of assigning depth values to pixels are as follows (see also Figure 5.5):

1. The pixel coordinates and depth values from the PMD camera are used for generating a tessellated depth surface (i.e. quad mesh) of the scene from this viewpoint.
2. The surface is transformed into the view space of the cameras. The intersections of view rays with the surface in this coordinate system define initial disparity values; the associated confidence values define the possible range. Thresholding the confidence values yields a set of valid and invalid depth coordinates in the quad mesh.
3. The areas of pixels associated to valid and invalid depth coordinates are thinned and serve as the initialization of a segmentation of the color image into depth continuous regions.

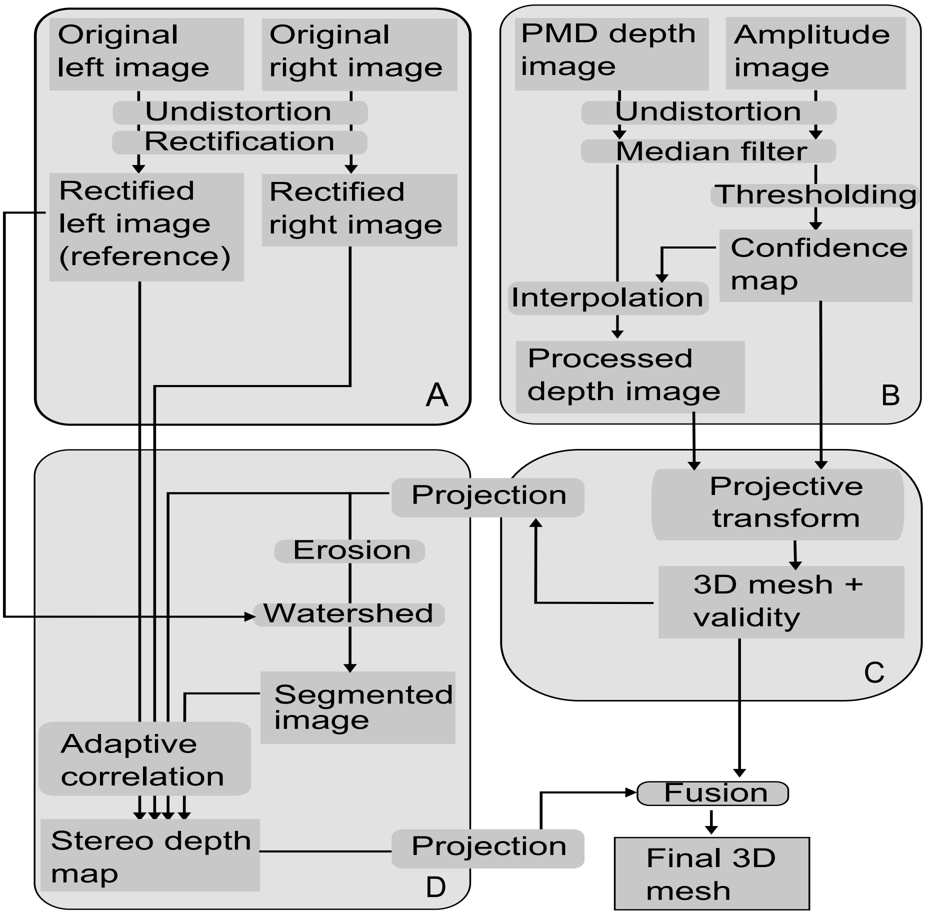


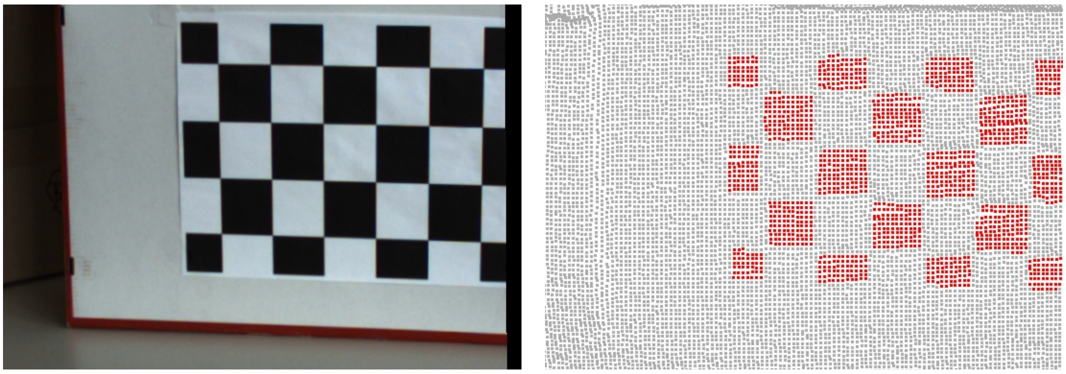
Figure 5.5: Algorithm overview including image processing steps for stereo (A) and PMD (B), 3D transformation of PMD data (C), stereo depth estimation (D) and fusion.

4. The segmentation steers adaptive windows for the correlation computation in a standard stereo algorithm, correcting the invalid depth coordinates of the surface mesh.

In the following, we discuss several details of these steps.

Mesh initialization and projection The intrinsic calibration of the PMD camera allows computing 3D coordinates from pixel location in the image plane and the corresponding depth value. For convenience, we connect the 3D coordinates to a piecewise bilinear mesh. The extrinsic calibration between the cameras allows transforming the mesh into the coordinate systems of the stereo camera. The intrinsic calibration of the stereo cameras (including a rectification) defines a projective transformation, which yields depth and confidence values per pixel.

The labels define two regions in the color images: a region of valid and a region



(a) Rectified Left View

(b) Stereo operation map

Figure 5.6: Color coded vertices of the surface mesh, where red vertices are invalid and will be corrected using stereo vision.

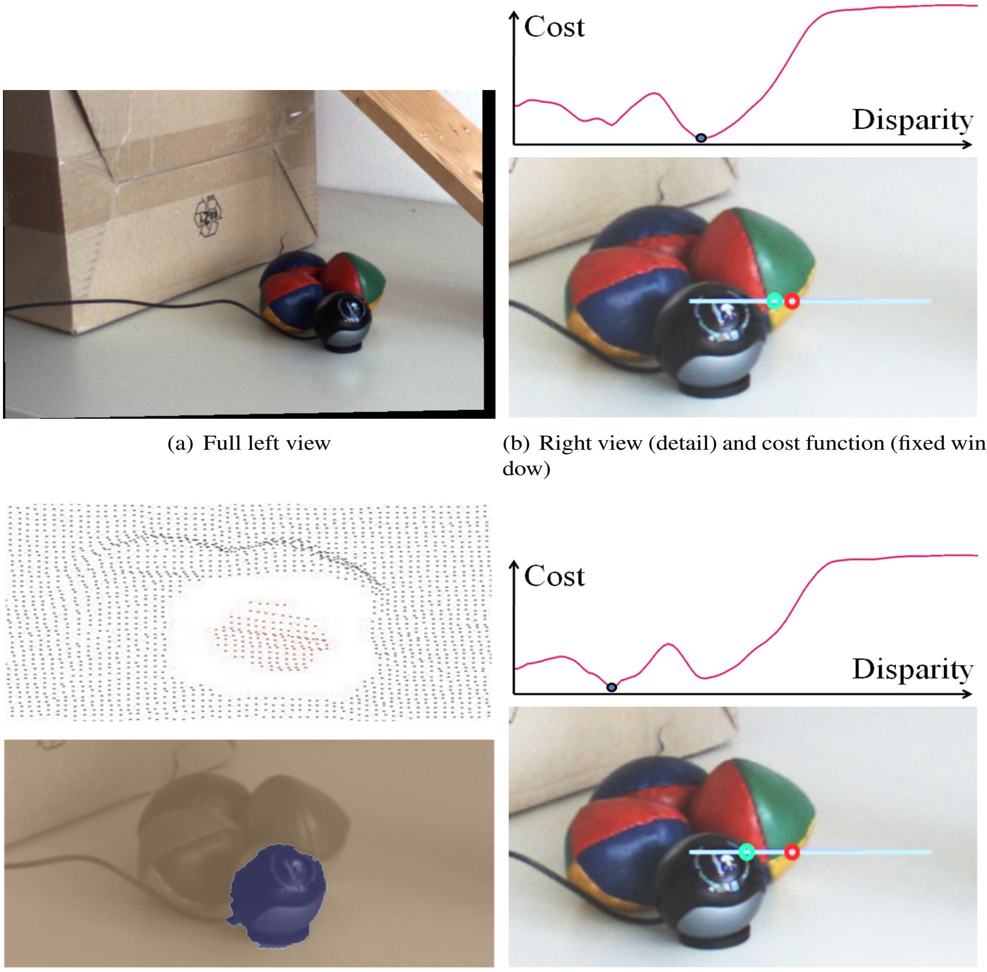
of invalid pixels, based on the binary confidence map. Figure 5.6 shows the projection of the mesh in the left camera. The projection of valid vertices are drawn as gray squares and invalid pixels are drawn as red squares.

On-demand stereo with adaptive windows Instead of computing a whole disparity map, the use of our stereo part is computing depth values only for vertices that are marked invalid. Furthermore, the projection of these vertices into the two rectified stereo views immediately yields an initial disparity guess.

An underlying assumption of correlation based stereo algorithms is that depth is unambiguous in the correlation window. This is not the case at depth discontinuities where objects may be occluded in only one of the views so that correlation of pixel colors fails to be a good indicator for correspondence (Figure 5.7(b) illustrates this situation).

A solution to this problem is to adapt the correspondence window to the (likely) object boundaries. Kanade and Okutomi [52] suggest to adapt the size and shape of the rectangular correlation window to local disparity characteristics. Boykov et al. generalize this variable window approach [9]. They compute for each pixel a new window. This window contains all pixels with an intensity close to the considered pixel. This way, they try to model the boundaries of the objects and the depth discontinuities. Hirschmüller [42] proposes a similar approach using multiple supporting windows. Unfortunately, all of these techniques are too costly to reach interactive rates at video resolution of 640 × 480 pixels.

Our main observation is that the object boundaries are only relevant if they are in regions whose depth values are labeled as invalid – otherwise the depth values have already been gathered based on ToF. Thus, we can use the information on valid and invalid regions for initializing a segmentation algorithm in the color images. The segmentation will then define the extent of the correlation windows



(a) Full left view

(b) Right view (detail) and cost function (fixed window)

(c) Eroded operation map and segmentation (both detail)

(d) Right view (detail) and cost function (adaptive window)

Figure 5.7: This figure compares correlation based stereo with fixed windows and with windows adapted to object boundaries computed from segmenting the color image into depth continuous regions. The whole scene is shown (a), while we focus on the group of balls and the webcam in front of the box (b-d). The eroded operation map is used to initialize the watershed segmentation (c) leading to a mask adapting the stereo correlation windows. The red circle shows the initial disparity guess and the green circle the disparity corresponding to minimum cost (b+d).

used in our adaptive window stereo algorithm. Exploiting the confidence information makes our approach both much faster and also more robust than only working with the color images.

From the many potential segmentation algorithms we use the *marker-controlled watershed* algorithm [99], which we have found to be robust while being fast enough for our application scenario. The idea is that valid and invalid regions serve as markers for the binary segmentation. Because of errors in the projections for vertices with incorrect depth (i.e. especially invalid vertices), color pixels are not necessarily labeled correctly. Consequently, the sets of valid and invalid pixels are eroded independently, leaving a set of unlabeled pixels in the proximity of object boundaries (see Figure 5.7(c)). These sets of valid and invalid pixels serve as the markers that initialize the segmentation as starting points. If objects have boundaries in the color images, the segmentation will accurately label pixels as being connected to the valid or invalid pixels. The resulting binary map restricts the correlation window.

Figure 5.7 shows the influence of this border correction filter: an object is too dark for the PMD camera, yielding wrong depth values and marked as invalid. A correlation based stereo algorithm with fixed window finds the wrong corresponding point (5.7(b)). After eroding the sets of invalid and valid pixels, the watershed algorithm segments the object along its boundary (5.7(c)). Restricting the window to the segmented object yields the correct correspondence (5.7(d)).

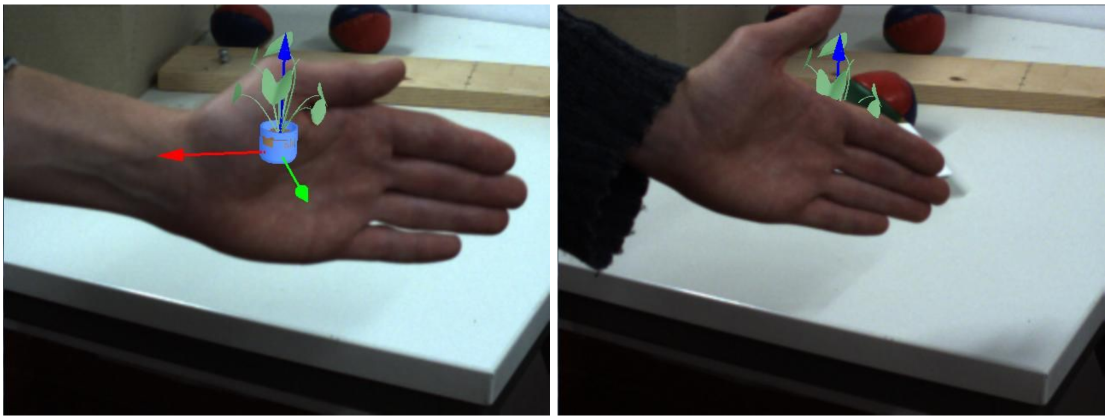
Final 3D mesh Finally, we replace the PMD depth of each invalid pixel with the computed stereo depth value. The resulting mesh is rendered into the z-buffer of one of the camera views using the color values from the camera. This allows using the depth values in interactive applications.

5.5 Results

As pointed out in Section 1.2, occlusion is one of the most fundamental factors in monoscopic depth perception, AR applications become much more immersive if occlusion handling is embedded. This is demonstrated in Fig. 5.8. Unlike model based techniques [10], our approach makes it possible to handle dynamic occlusions like a hand in front of the virtual objects. This is very practicable as most AR application focus on manual interaction of virtual and real objects.

Similar approaches for occlusion handling in AR exist. While Kanbara et al. [53] have utilized stereo vision, Fischer et al. [19] also use ToF depth sensing. We have compared the raw output of the PMD camera projected into the view of the color image and the raw stereo data with the results of our approach (see Fig. 5.9). A dark object in this scene is assigned incorrect depth values delivered by the PMD camera and, consequently, is mapped to the background through the interpolation process (Fig. 5.9(c)). Fig. 5.9(d) shows the depth map obtained using correlation based stereo with a fixed window as used in other stereo algorithms aiming at interactive frame rates. The images in 5.9(e) and 5.9(f) show the improved results using our algorithm. Notice the appearance of dark objects while they are not captured at all by the PMD camera.

Several parameters influence the performance of our system: the frame rates we obtain are limited by the cameras and the transmission to roughly 11 fps for



(a) no occlusion

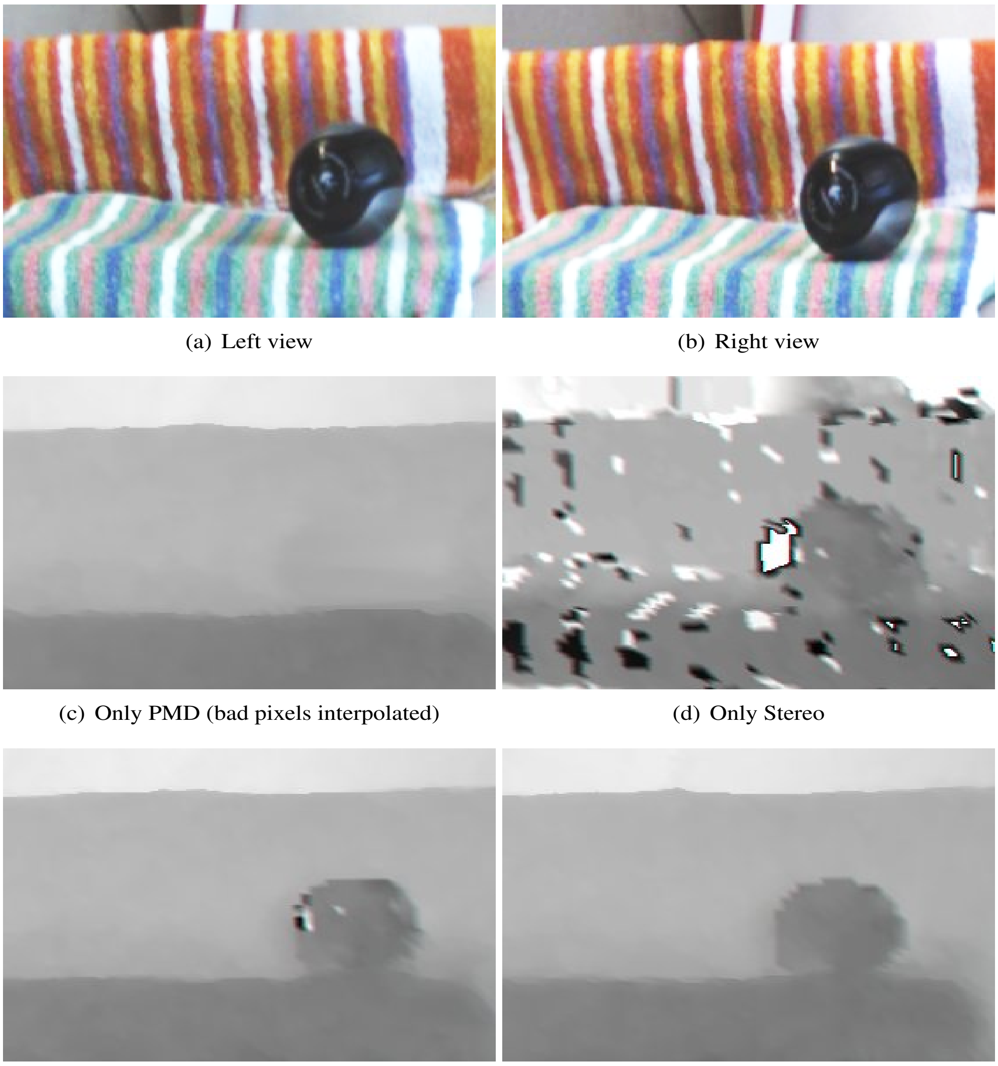
(b) occlusion

Figure 5.8: Occlusion strongly enhances the depth impression of the scene.

the C++ implementation on our test system, an AMD Athlon 2.0 GHz Dual Core Processor with 1 GB RAM. The additional cost of our algorithm depends on the size of the correlation window and the number of depth values that have to be corrected using stereo on-demand. Table 5.1 compares several situations, where we have chosen the thresholds so that approximately 250 resp. 500 depth values were considered invalid. The computation times clearly show that there is a linearity between time and correlation window size for stereo computation on the one hand and on the other hand, the ratio between the amount of corrected pixels and computational timings is linear as well. We therefore expect our system to be capable for real-time applications with even higher numbers of invalid and hence corrected pixels using recent hardware.

Acquisition	91 ms		
Preprocessing	23 ms		
Stereo+PMD	Window	≈ 250 px	≈ 500 px
	5×5	**9.5** ms	20.5 ms
	7×7	16 ms	33.3 ms
	9×9	25 ms	55 ms
	11×11	42 ms	88 ms

Table 5.1: Performance on an AMD Athlon 2.0 GHz Dual Core Processor with 1 GB RAM. Window is the size of the correlation window used for stereo. The computation times are compared for each step while capturing a scene with approx. 250 and 500 corrected pixels.



(a) Left view

(b) Right view

(c) Only PMD (bad pixels interpolated)

(d) Only Stereo

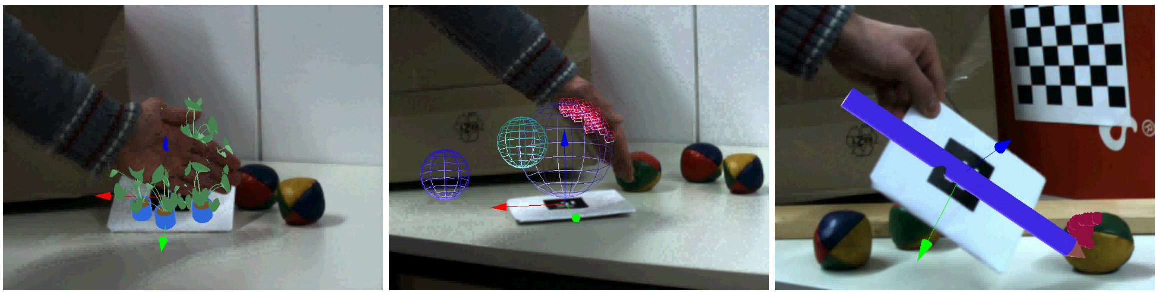
(e) Stereo and PMD data

(f) plus adaptive windows

Figure 5.9: We compare the depth map acquired with the PMD camera (c) and using correlation based stereo (d) with our results, i.e. correcting low confidence areas of the PMD range image using stereo on-demand with a fixed window approach (e) and using adaptive correlation windows based on segmentation (f).

5.6 Discussion

Our system provides a framework for interactive AR applications, where the depth map is necessary for visual or physical interaction between synthetic and real objects. Our approach is using low cost, light weight components and exploits the properties of both sensor types. In particular, we use range data for narrowing the disparity search and adapting the correlation windows to potential depth disconti-



(a) occlusion (b) collision (c) painting

Figure 5.10: Exemplary interaction concepts for our system.

nuities.

We demonstrate that the resulting system produces reliable depth information that can be used for handling occlusions. The depth data can also be used for collision detection and other interactive solutions. Some of the possible interactions are demonstrated in Figure 5.10. The dynamic and global depth map of the scene would also allow computing shadows for virtual objects, or using higher quality rendering techniques for further improving the realism of virtual objects [8].

As an alternative approach, it would be interesting to apply different segmentation algorithms for adapting the stereo windows. Feris et al. [18] use region growing in a similar situation. Finding the best balance between performance and accuracy in this step is important future work. In addition, it might be possible to make use of other information than confidence and color, such as range data or predictions from preceding frames. Another step in enhancing the algorithm would be to define a confidence measure for the stereo data and use it for further controlling the depth reconstruction. This could reduce the error in regions where both systems fail, for example large dark and homogeneous objects.

Our system can be improved in several aspects. The synchronization of the PMD[vision] 19k camera as part of the software system results in a maximum of 11 fps – using a hardware solution would allow exploiting the maximum frame rate of the cameras. The computations necessary for fusing the data and improving the depth images are easy to distribute to several cores so that exploiting a higher input frame rate would be easy if the system were coupled with a modern central processing unit (CPU). As mentioned in Section 5.2, the accuracy is depending on the PMD range data, and we think that this makes it important to develop approaches that lead to an increased working range.

In order to increase the working range, the concept of confidence measures can be exploited to fuse several depth images from the PMD camera alone. Hereby, the integration time is varied as it is done in HDR imaging approaches. We will describe our idea and experimental results in the following chapter which has been previously published as [39].

Chapter 6

Exposure Fusion for Time-of-Flight Imaging

6.1 Introduction

As pointed out in the course of this thesis, ToF based depth sensing is utilized in more and more fields of application, among others simultaneous localization and mapping (SLAM), motion capturing in gaming and pedestrian detection in automobiles [79, 125, 88, 61]. All of these applications need the sensor to capture reliable depth data within a range of several meters.

We described in detail in Chapter 3 how PMD cameras are operated. One important parameter that has to be defined for each measurement is the integration time. The integration time indicates how long the sensor is exposed to accumulate photons from correlation signal samples in order to calculate the desired phase shift. A deficiency in numbers of photons leads to uncertain results dominated by noise. On the other hand, too many samples potentially lead to saturation and photons are no longer counted, which also results in errors. Therefore, setting the correct integration time is crucial for correctly measuring the distance. Usually, the operator of the device has to define the integration time for the sensor manually. May et al. [79] have proposed a control mechanism that adjusts the integration time during operation. Here, the integration time is set by means of a feedback controller that assumes that the integration time is optimal when the mean intensity of the captured image is at a pre-defined ideal value.

Nevertheless, such an auto-exposure mode is only able to find one globally optimal exposure time for one scene. And just as in regular photography, the resulting image might still have under- and over-exposed regions. As the depth measurement relies on the reflection of an emitted light signal, under- and over-exposed regions lead to errors in depth estimation, as explained above. In fact, each pixel has its own optimal integration or exposure time, which *unlike traditional photography* depends on the *distance and reflectivity* of the captured object itself.

In this chapter, we adapt recent solutions from computational photography to this problem. Instead of optimizing for a global integration time, we capture several images and search locally in each exposure for regions which provide most accurate distance data. This poses two challenges: first, the sources of error in the sensor are different from traditional over- and under-exposure in imaging. Second, ToF depth sensing is useful mostly in real-time applications meaning the solution has to be computed in fractions of a second.

We implemented a method for capturing ToF range maps in order to provide high quality depth data for the full theoretic range of the camera. Our approach is inspired by HDR imaging, but faces the challenge of dealing with depth instead of color information. Therefore, we propose new measures for the quality of the depth data locally in the original images that lead an image fusion process. These measures are both inspired by a similar approach for color images by Mertens et al. [80] and founded on research in image quality measures [77, 31]. The ToF camera we use returns amplitude images that refer to the correlation signal strength that has been captured by the sensor. We use only these images and the distance data to compute our quality measures. Hence, our solution does not need any calibration process in order to enhance the reliability of the depth images. We implemented a real-time solution that exploits the capability of the PMD[vision] CamCube 3.0 camera to capture four images with varying integration times almost at once.

In order to demonstrate the superior quality of the fused depth maps, we captured known planar objects at varying depths and measured the error from the residuals of a least-square plane fitting in the planar regions of the image. Our results show that the fused data is more accurate than data from the ideal exposure time even for a single planar depth region and gives much lower errors if several planar regions at varying depths are taken into account.

We further apply our fused range maps for point cloud alignment and compare the results of the 3D reconstruction of indoor environments with them produced by single exposure depth images.

6.2 Related approaches

To our knowledge, this is the first approach to combine several exposures with varying integration times of a ToF camera in order to enhance the quality of the depth maps. In the following, we relate our work to approaches concerning the integration time in ToF imaging as well as alternative approaches to enhance the dynamic range of the depth sensing.

Hardware First, the camera manufacturers try to enhance the dynamic range as much as possible. MESA Imaging, producer of the Swissranger™ cameras has developed a solution that allows to control the integration time per pixel individually [11]. Such an enhanced pixel stops the integration as soon as the capacitance exceeds a pre-defined threshold. Unfortunately, the power consumption of such a pixel enhancement is too high in practice and additionally, the used integration time for each pixel has to be stored and transferred in order to reconstruct a homogeneous intensity image. This lead MESA Imaging to not implement such a feature in their products.

There are ambitions to extend the range for ToF imaging by PMDTec as they are offering a plugin to enable modulation frequencies down to 1 MHz. This enhances the working range theoretically up to 150 m. However in practice, the accuracy would be strongly reduced and the illumination unit has to be amplified as well. Nevertheless PMDTec promises an extended range of 30 m.

As pointed out in Section 3.2 on page 28, there are numerous approaches that deal with denoising the distance data captured by ToF sensors. While most of these approaches either use additional sensors [45, 75, 130] or rely on elaborately generated calibration data [51, 71, 76], our approach can be applied to all ToF cameras that deliver amplitude and distance data without any preparation of the sensor. While receiving very good results, Lindner and Kolb [71] use an additional camera and pre-captured calibration data in order to correct the error resulting from differently reflecting objects.

Software Similar to our approach, Schuon et al. [106] presented a method based on super resolution. Here, several noisy depth maps captured from slightly different positions are combined to one high resolution, high quality depth map. While this approach has been successfully extended and applied to 3D shape scanning [107, 14], it does not provide – in contrast to our approach – the enhanced depth imaging in real-time.

6.3 Motivation

As already mentioned, our approach is inspired by HDR imaging. Here, several exposures of the same scene are captured with varying exposure times. This leads to images with a varying amount of details in different regions of the image. These images are fused together in order to keep the details visible in all image regions. We refer to the book of Reinhard et al. [96] for a complete overview in HDR imaging.

Usually, the different images are aligned to one HDR radiance map which can not be displayed without specialized devices [15]. It has to be transformed back to low dynamic range by tone mapping to enable the visualization on a regular display. This process has been shortened by Mertens et al. [80]. Thereby, the images are fused directly into a single low dynamic range image that contains all the details from a collection of differently exposed images. For each image pixel a weight is calculated and the final result is an affine combination of the images.

The fusion of color images has already been realized by Goshtasby [31]. He proposes a measure for the entropy of each image pixel and fuses the images based on this measure in a gradient-ascent approach which is not utilizable in real-time applications. In contrast to this, Mertens et al. [80] aim on the same outcome of fusing images with varying exposure times. They propose three quality measures and merge the images in a fast pyramid based algorithm.

These quality measures are not suitable for range maps in general. We therefore adapt the quality measures to the characteristics of ToF range images. The measure should define a confidence value of the depth data as applied in many other approaches [78, 61]. While Frank et al. [22] show that the amplitude value is an optimal indicator for the confidence of the range data, Reynolds et al. [98] demonstrate that a trained random forest outperforms simple amplitude based thresholding mechanisms. Apart from that Foix et al. [20] recently presented an approach that models the uncertainty only from the depth data. Based on these inconsistencies in the literature we develop and evaluate new measures. We explain our choices in the next section.

6.4 Algorithm

Fusion process

Our fusion algorithm is similar to the one described by Mertens et al. [80] which has been successfully used in recent real-time applications [1]. We take several exposures of a scene and fuse the depth images together. Each exposure is multiplied per pixel with a weight map W_k where $k = 1, \ldots, N$ indicates the number of exposures. This weight map is constructed as an affine combination of several individual quality measures. We define

$$W_k = M_C^{w_C} \times M_W^{w_W} \times M_S^{w_S} \times M_E^{w_E}$$

with the quality measures M (resp. **C**ontrast, **W**ell-exposedness, **S**urface and **E**ntropy) and $\times$ denotes a per pixel multiplication. This multiplication leads to fact that each quality measure is able to limit the influence of an exposure by setting a pixel to zero. Each quality measure M is weighted with an corresponding exponent $w \in \{0, 1\}$. The weight map is normalized so that the weight of all exposures k sums up to one for each pixel.

The fusion is realized as a multiresolution blending. Instead of fusing directly the full resolution depth maps, image pyramids are computed and fused as proposed by Burt and Adelson [12]. The resulting fused depth map R can be reconstructed from the Laplacian pyramid $\mathbf{L}\{R\}$. The l-th level is defined by

$$\mathbf{L}\{R\}^l = \sum_{k=0}^{N} \mathbf{G}\{W_k\}^l \times \mathbf{L}\{D_k\}^l,$$

where D_k denotes the depth map from the k-th exposure. Each level l of the pyramid is constructed by a weighted sum of the corresponding levels of a Laplacian pyramid over all exposures. The weights are obtained from the l-th level of the Gaussian pyramid of the weight maps. See Figure 6.1 for a schematic overview about the process. Note that this fusion scheme slightly enhances the quality, however it can also be replaced by a simpler full resolution blending.

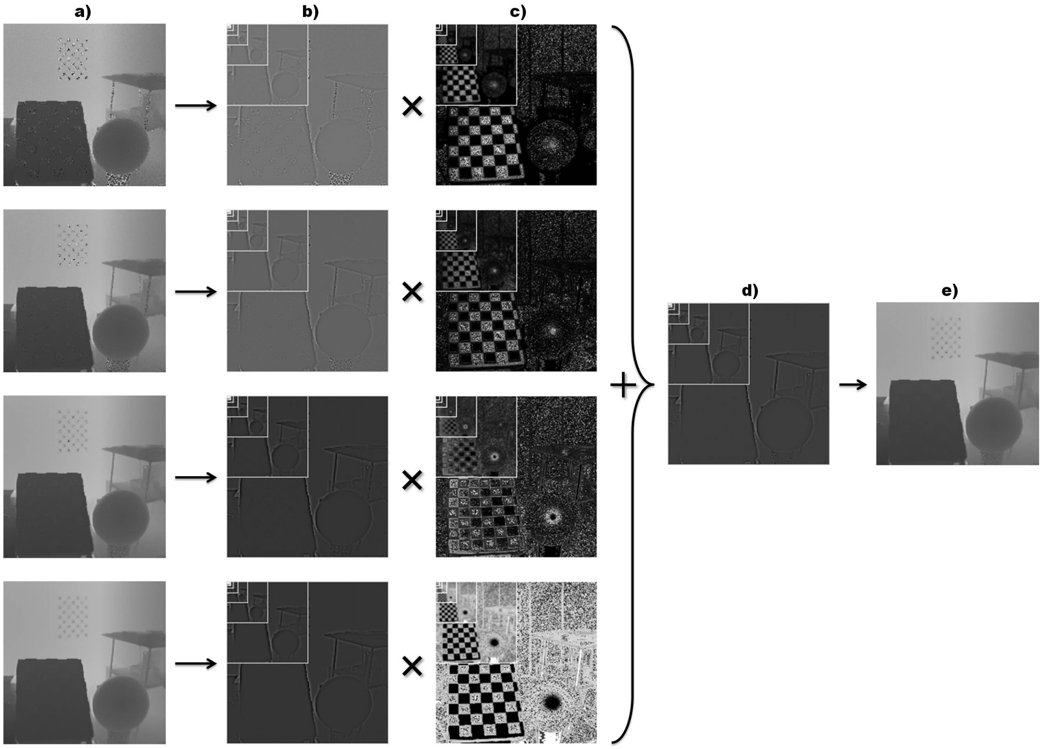


Figure 6.1: Exposure fusion principle: a) Captured depth maps, b) Depth map – Laplacian pyramids, c) Weight map – Gaussian pyramid, d) Fused pyramid, e) Final depth map (after [80])

Quality measures

In the following, we describe the definition of new quality measures for the depth map fusion in detail. Note that these measures are not entirely calculated from the depth images, but also based on the amplitude images. We enumerate the image pixel indices as i and j. The distance image D with distance values D_{ij} is normalized to $[0,1]$ by setting a linear mapping. Distances with the theoretical maximal distance of 7.5 m are mapped to 1, while a distance of 0 m is mapped to zero. The amplitude image A with amplitudes A_{ij} is also normalized with one global value that is mapped to 1. Note that the amplitude does not have a theoretical maximum. It is bounded by the technical properties of the chip, hence we bound the maximum at a value where the sensor is not yet saturated.

Contrast M_C As pointed out in Section 3.2 and illustrated in Figure 3.8 on page 35, one big issue with ToF depth images are so called *flying pixels*. Due to aliasing effects, the distance along depth discontinuities is computed from photons collected by the sensor from foreground and background. This leads to wrong distances that do not necessarily lie between the values of fore- and background. However, the amplitudes – which are also measured per pixel – along the depth discontinuities do lie between the fore- and background values. Hence, we can define a quality measure that fosters image regions where the depth discontinuities do not lead to flying pixels. These regions are identified by the contrast in the amplitude image which leads us to define

$$M_C = \|\Delta A\|.$$

We apply a 3 × 3 Laplacian filter to the amplitudes images and use the absolute values of the filter response which yields an indicator for contrast. In the amplitude images a strong contrast occurs usually along depth discontinuities as the reflectance of the foreground object differs from the object behind. Thid contrast is assumed to be stronger than the contrast due to differently reflecting textures. Note that this measure has also been used by Mertens et al. [80] in order to enhance the contrast in the resulting image.

Well-exposedness M_W For ToF cameras the amplitude image indicates under- or overexposure, hence the amplitude can be used as a confidence measure. As already mentioned, we normalize the amplitude images. We determine amplitude values A_{min} and A_{max} for under- and overexposure and map all the values in between linearly to the interval $[0,1]$. All values outside this range are mapped to zero or one respectively. We calculate each pixel W_{ij} of this quality measure M_W as

$$W_{ij} = e^{\frac{-(A_{ij}-\alpha)^2}{2\sigma^2}}$$

with $\alpha = 0.5$ and $\sigma = 0.2$. We adapt this quality measure from the so-called well-exposedness measure from Mertens et al. [80]. They argue that intensities close to zero indicate underexposure and close to one overexposure respectively. In our adaption the pixels with an optimal normalized amplitude value of 0.5 get the highest weighting. Note that the critical part is the determination of A_{min} and A_{max}. They can either be obtained from the camera manufacturer or by capturing a wall from a fixed distance. Then plot the mean distance and amplitude values while varying integration time from low to high. The mean distance will change drastically as soon as the sensor is under- or overexposed. From this boundaries A_{min} and A_{max} can be determined.

Surface M_S Besides these two quality measures that already lead to promising results, we defined a further one based on the measure of the structural similarity [116] and its adaption to range maps [77]. A measure for the surface roughness can be defined as

$$M_S = 1 - \frac{(\sigma - \mu^2)}{max(\sigma - \mu^2)}$$

where σ is the Gaussian filtered version of D^2, while μ is a Gaussian filtered version of D. The difference $\sigma - \mu^2$ correlates with the frequencies in the images. Its value is high for high frequencies and vice versa. These high frequencies are an indicator for noise in the depth image as the captured scene is assumed to be smooth in relation to camera noise. This difference is divided by its maximal value and subtracted from one so that the measure M_S is high in smooth regions. The smoothness indicates the absence of noise and leads to the assumption that the depth values are correct.

Entropy M_E Similar to the approach from the well-exposedness measure M_W, Goshtasby [31] describes the idea for image fusion to strengthen image regions that contain the most information. A measure for the amount of information is entropy. The entropy measure M_E contains an entropy value E_{ij} for each pixel. It is calculated for a local histogram around each pixel of the image as

$$E_{ij} = -\sum p(A_{ij}) \log_2(p(A_{ij}))$$

where p contains the histogram counts from a 9×9 neighborhood of each pixel and each histogram contains 256 bins. The entropy of an image has the property that it only depends on the image histogram. This leads to a disadvantage of the entropy. The image information is defined as uncertainty which is maximal for noisy images.

See Figure 6.2 for a side by side comparison of all quality measures for an example scene.

Discussion

We further have to define the number of exposures k that we will use for image fusion. Our algorithm works with an independent number of exposures. In our real-time implementation, we fuse four images because the PMD[vision] CamCube 3.0 provides a capture mode that allows to take four successive frames without transferring data in between. Due to the short integration times about max. 5–10 ms, these four exposures do not differ significantly even for scenes containing motion. Furthermore, the computation costs fusing four images still allow interactive frame rates.

In order to correct the distance values in real-time, the quality measures have to be computed efficiently. We use a profiler for determining the attended computation time of each measure. See Figure 6.3 for an illustration of the profiling

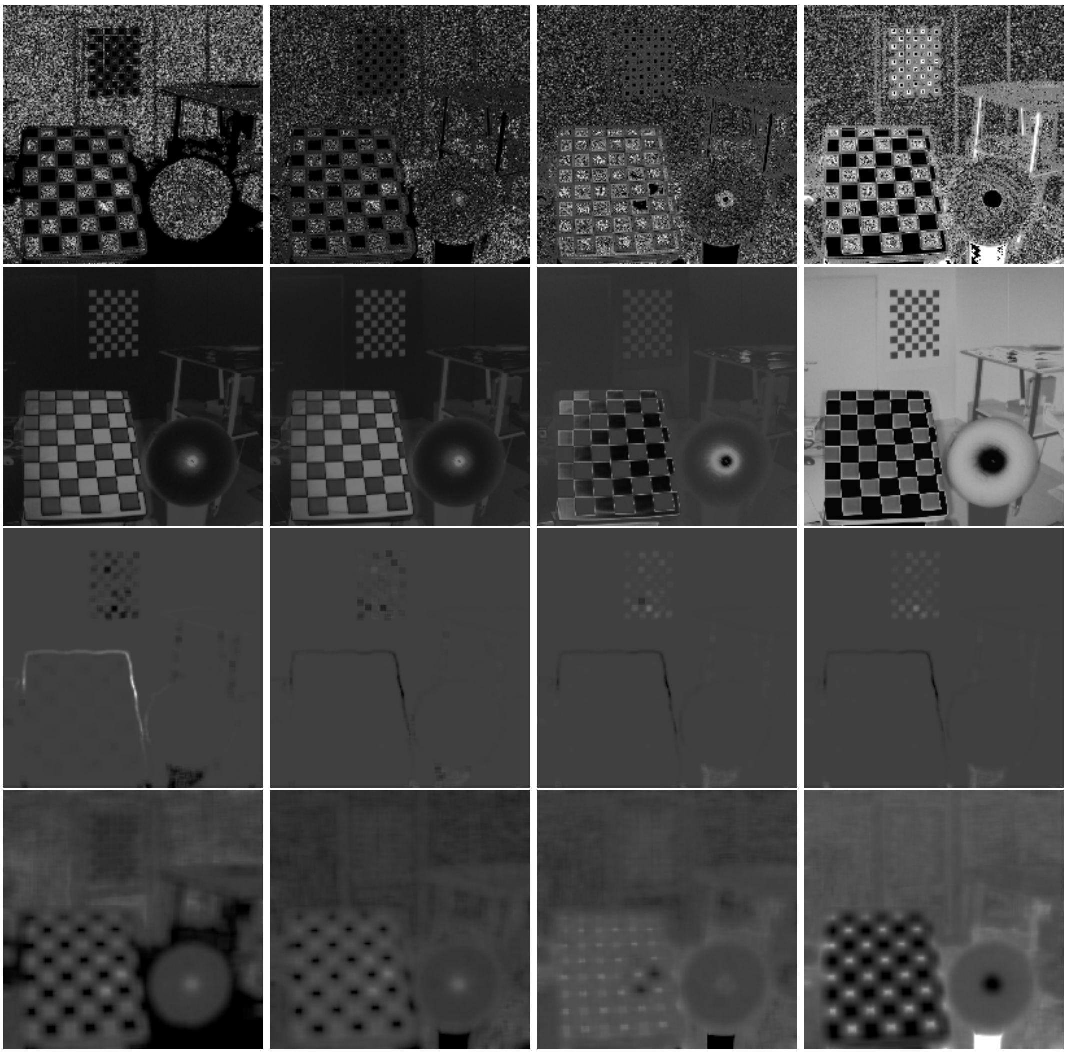


Figure 6.2: Comparison of all weights side by side. Each column is one integration time. Rows are sorted from top to bottom showing M_C, M_W, M_S and M_E.

data measured during live capture and display of the fused depth images. The blue bars in the front row display the computation time contribution during the capture, fusion and anything else (e.g. rendering). The red bars are the weighting functions inside the fusion algorithm. This figure clearly shows that the entropy calculation is the bottle neck in our implementation. It leads to a decrease in frame rate in the real-time implementation down to 2–4 fps even if optimized algorithms are used for the calculation of the logarithm [115] and the number of unnecessary additions in the summation is minimized. Without calculating the entropy the whole fusion process is computed in 0.046 seconds on standard laptop computer with a dual core processor running at 2.16 GHz and equipped with 2 GB of RAM.

We therefore further evaluate our algorithm to identify the impact of each qual-

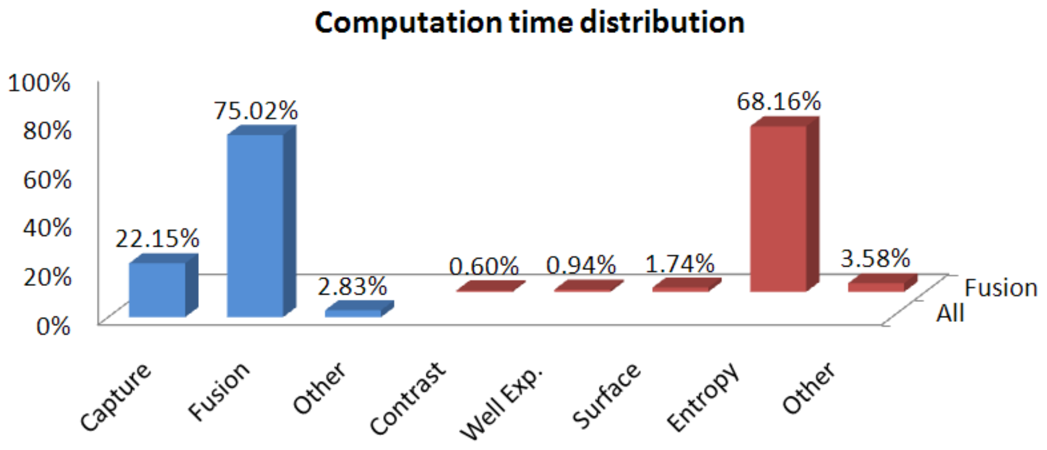


Figure 6.3: Results from profiling: blue bars show the distribution for the complete program, red bars the distribution inside the fusion algorithm.

ity measure on the accuracy of the depth maps.

6.5 Evaluation

We implement several test environments in order to demonstrate the superior quality of the fused depth maps over depth maps acquired with a single integration time.

Reference

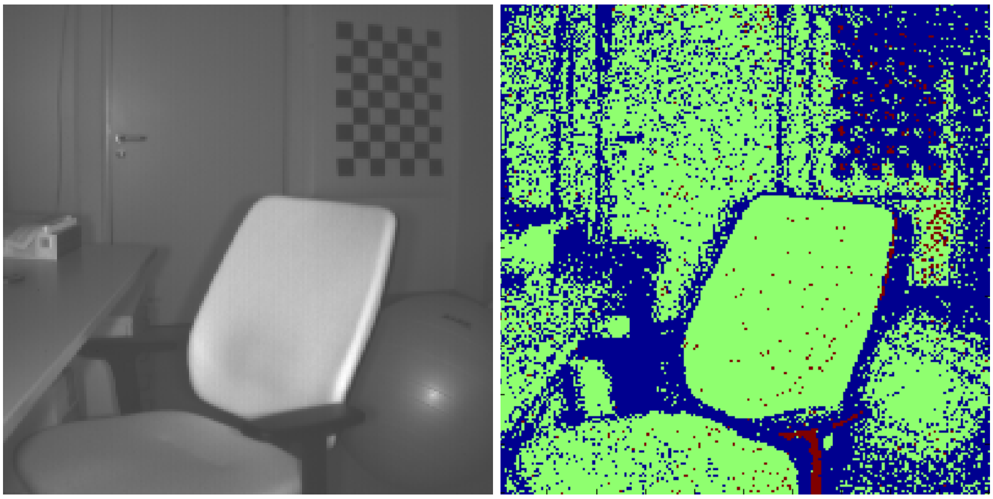
First, we need to define the integration times for our test. Therefore, we implemented a proportional-integral-derivative (PID) controller [122] approach following the ideas of May et al. [79]. The integration time is set so that the mean intensity of the captured image is optimal. In the following, we refer to this integration time as the ideal integration time t'. In addition, the number of exposures and the integration times for each exposure have to be defined for our test cases. We decided to use the following scheme for the first test:

$$t_i = 2^{i-1-\frac{N}{2}} t'$$

where N is the number of exposures and $i = 1, \ldots, N$ indicates the exposures.

Robustness test

As a first comparison between the fused images and images captured with the ideal integration time, we analyzed the standard deviation of the distance values of a



(a) PMD[vision] CamCube 3.0 intensity image

(b) Comparison of standard deviations for single exposure and fused solution

Figure 6.4: Static scene used in the robustness test.

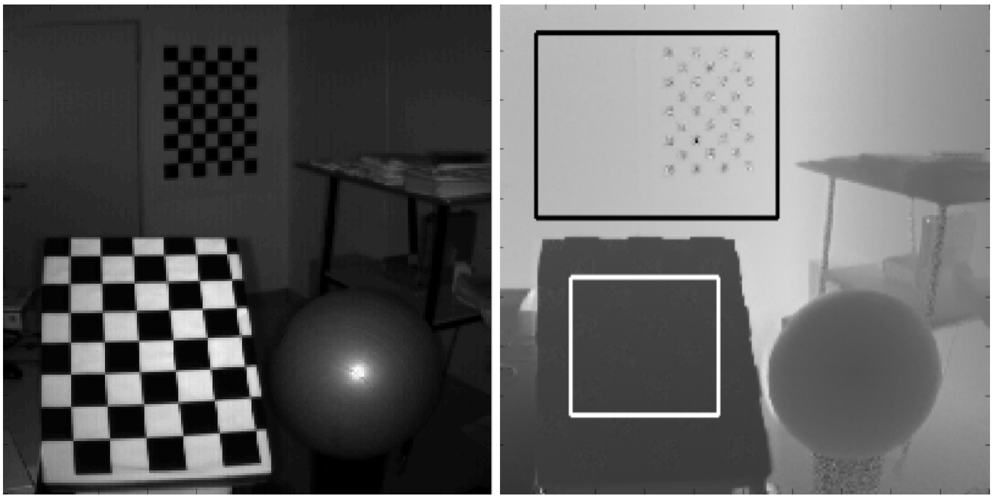
static scene over a time period of 50 frames. See Figure 6.4 for an intensity image of the scene (left). We further compare the standard deviations per pixel (right) – we colorize the pixels depending whether the standard deviation is smaller for the single integration time exposures (red) or for the fused solution (blue). All pixels where the standard deviation does not differ more than 5 mm are marked green. This illustrates that the fused values are more stable especially around depth discontinuities and in textured regions. The mean standard deviation over all pixels in the fused images over time is reduced significantly from 2.94 cm to 2.17 cm.

Plane fitting error

In a second test, we place two planar objects (see the checkerboards in Figure 6.5) at different depths in the scene. We select these two regions of interest (ROI) manually. Note that the far checkerboard is mounted on the wall. We captured a series of four exposure with the same scheme as in the stability test case.

As our algorithm enhances directly the depth maps, we have to transform the distance values of each pixel in both ROIs (near and far) into 3D coordinates. We use the fixed intrinsic parameters of the PMD[vision] CamCube 3.0 to calculate the 3D points for each pixel of our fused depth maps as well as to compute a point cloud from the depth map obtained with by a single exposure.

We then fit a plane into each ROI in 3D coordinates using a principal component analysis and check the correct rough orientation of this planes normal. The perpendicular distance from each point to the plane is minimized. The error is defined as this distance. We then compare the mean squared error (MSE) for this



(a) PMD[vision] CamCube 3.0 intensity image

(b) Fused depth map with two ROIs (white: near, black: far)

Figure 6.5: First plane fitting test scene.

plane fit and the computation time over various weighting combinations. For each ROI we compute the MSE for the data captured with the optimal global integration time and for the fusion results using all possible combinations for the weighting exponents. In addition, we measure the computation time for the fusion process (see Table 6.1). The first row contains the results for a single exposure with the optimal integration time determined by the algorithm of May et al. [79]. The second rightmost column contains a weighted sum of both MSE. We use the ratio between the MSE from the single exposure as weight for the sum, so that both MSE are equal. The rightmost column shows the error in relation to the weighted sum MSE from the single exposure – values below one indicate an enhancement.

The primary outcome is that the best weighting combination in this setup is a combination of all presented quality measures (see the last row). Our results show that the error is reduced by about nearly 38%.

In order to stress the positive effect of our weights we further compare our solution with a simple approach. We use the normalized amplitude image directly as weight. This results in a small error for the near plane, however the far plane fitting leads to an MSE of 0.02346 what is even larger than in the single exposure. We show the (far) plane fitting results in Figure 6.6.

We illustrate the correlation of error and computation time in Figure 6.7. The plot displays the time and error for the combinations of the three most important quality measures – we left out the surface measure M_S for clarity, because it has no effect in the planar regions, however weights correctly around depth discontinuities (see Figure 6.2 on page 70). The contrast measure M_C can clearly be identified as the most effective measure. Adding the well exposedness measure M_W slightly reduces the error. The entropy measure M_E further reduces the error,

M_C	M_W	M_S	M_E	**Timing**	**MSE (near)**	**MSE (far)**	**Sum (MSE)**	**Error**
0	*0*	*0*	*0*	*0.000*	*0.000761*	*0.0223*	*0.0446*	*1.000*
0	0	0	1	1.248	0.001033	0.0215	0.0518	1.160
0	0	1	0	0.478	0.001689	0.0199	0.0695	1.557
0	0	1	1	1.304	0.001010	0.0168	0.0465	1.041
0	1	0	0	0.452	0.001287	0.0256	0.0634	1.421
0	1	0	1	1.263	0.000786	0.0206	0.0436	0.977
0	1	1	0	0.507	0.001254	0.0190	0.0558	1.251
0	1	1	1	1.317	0.000769	0.0161	0.0386	0.866
1	0	0	0	0.466	0.000414	0.0178	0.0299	0.671
1	0	0	1	1.275	0.000401	0.0175	0.0293	0.657
1	0	1	0	0.504	0.000413	0.0176	0.0297	0.666
1	0	1	1	1.324	0.000400	0.0174	0.0292	0.653
1	1	0	0	0.476	0.000383	0.0171	0.0283	0.634
1	1	0	1	1.296	0.000375	0.0169	0.0279	0.625
1	1	1	0	0.537	0.000383	0.0169	0.0282	0.631
1	**1**	**1**	**1**	**1.358**	**0.000375**	**0.0168**	**0.0278**	**0.623**

Table 6.1: Overview about the effect of each weight on accuracy and computation costs.

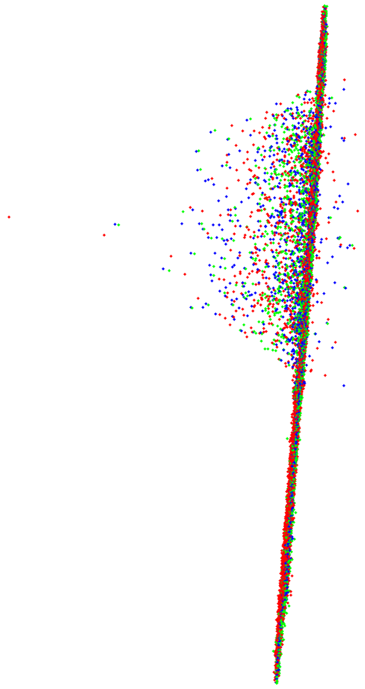


Figure 6.6: Close up on the far plane: green dots indicate measures from our fused results, red the single exposure and blue the simple amplitude weighting.

however at the expense of the computation time. Note that these timings are from the MATLAB implementation, however they confirm the trend from the profiling results of the real-time C++ implementation from Figure 6.3. Nevertheless the entropy is a suitable quality measure if the computation time is irrelevant.

We did the plane fitting test on further example scenes. Figure 6.8 shows two walls in an indoor environment. The wall on the right is close to the camera and untextured while the facing wall is textured. Note that the depth variance of the walls is far below the precision of the camera and we can hence assume planarity. We again fit a plane as in the previous example. We compare our fusion result with two simpler approaches. First, we did not apply the multiresolution fusion scheme, but simply computed the *weighted sum* for each pixel. Second, instead of using our derived quality measures, we again use the *normalized amplitude* directly as weight. Besides calculating the MSE, we further compare the estimated angle between the two walls which should be 90° (see Table 6.2). We included the results from the best single exposure of the sequence.

Method	MSE (face)	MSE (right)	Angle
Multiresolution	**0.07888**	**0.067636**	91.57
Weighted sum	0.07912	0.067819	91.61
Amplitude	0.07895	0.067769	91.60
Single exposure	0.07893	0.067643	**91.48**

Table 6.2: Error values from second plane fitting test scene for comparison with simpler weighting and fusion schemes.

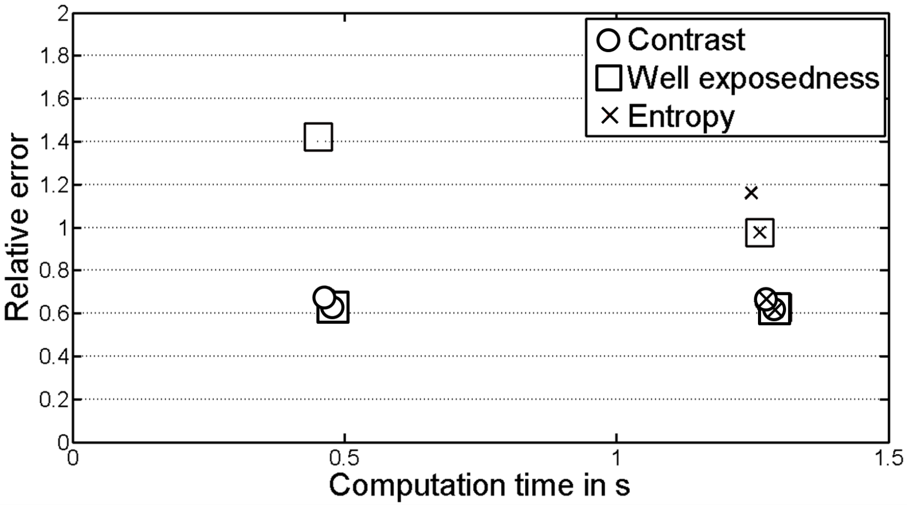
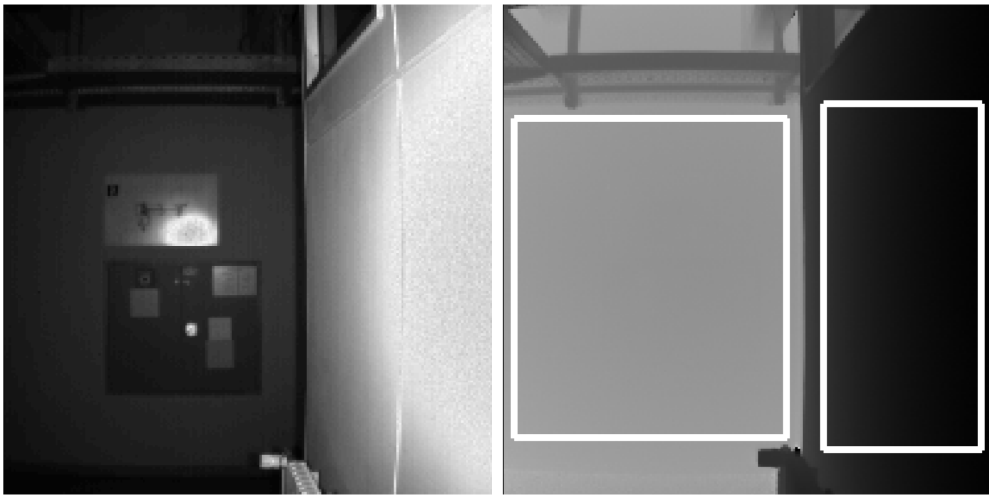


Figure 6.7: Plot of error versus timing for various weighting combinations.



(a) PMD[vision] CamCube 3.0 intensity image

(b) Fused depth map with two ROI (face and right)

Figure 6.8: Second plane fitting test scene.

3D reconstruction

A potential area for the usage of ToF data are autonomous robots and the SLAM algorithm. In order to determine the position of the robot (and hence the camera) the captured depth maps have to be registered. One way of registering is the alignment of point clouds. We therefore evaluate our method by performing a point cloud alignment by means of the well-known ICP algorithm [7, 13]. We then compare the alignment error and the convergence.

In our experiment, we mount the camera on a professional tripod and capture a static indoor scene by rotating the tripod stepwise. We use six exposures for fusion, then rotate the tripod by 10° and capture another series of exposures. For each position, we compute two point clouds. One directly from a single exposure with an optimal integration time, the second from the fused depth map. This results in two pairs of point clouds that have to be aligned. We use the ICP implementation from Kjer and Wilm [59] to determine a rigid transformation. For our fused solution the algorithm converges equally fast but the final error is smaller (see Figure 6.9).

Further we compared the resulting transformation with our manually defined ground truth – a rotation by 10° around the y-axis. We define the rotation error as the deviation from the identity

$$e(R, R_1) = ||I - RR_1^T||_F,$$

where R is the assumed correct rotation and R_1 the one we test, while $||\bullet||_F$ denotes the Frobenius norm. In contrast to a simpler error measure like the deviation in the rotation angle, this error also provides a measure for deviations in the other axes. Our fused solution results in an error of 0.0956 while the single exposure solution produces an error of 0.1598. This is an reduction of about 40%.

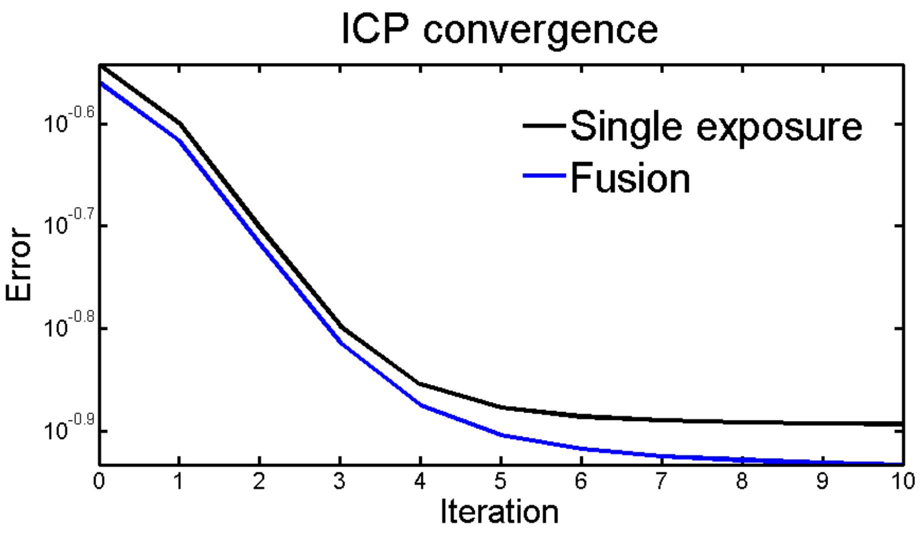


Figure 6.9: Convergence of the ICP algorithm for single exposure depth maps and our fused depth maps.

Limitations

Besides the shown positive examples our method is also limited. In some scenarios the fused depth maps are of equal quality as a single exposure. Table 6.2 shows that our method is not always far better than simpler approaches. This is the case if all objects have good (Lambertian) reflection properties and their distance is in a limited range. Our method works best if the distances and reflection properties are highly varying. However, neither strong noise in certain areas that are further than the theoretic limit of 7.5 m, nor severe over-saturation can be resolved properly. In addition, it is necessary that none of the input images is completely noisy.

6.6 Conclusion

We have presented and evaluated a new method to enhance the performance of ToF imaging devices. We developed test methods that do not need any extra hardware like a laser scanner in order to estimate the quality of our method. Our method successfully fuses several exposures into a single depth map and is on one hand not limited in the number of exposures and on the other hand fast enough to perform in real-time.

Our method not only works for fusing depth maps captured with different integration times, it also allows the combination of images with other varying parameters like the modulation frequency. We expect our method to achieve even better results for future ToF cameras with an extended theoretic range.

Chapter 7

Conclusions

We conclude this thesis with a summary followed by a short discussion about the possible impact of our approaches and also the limitations of our solutions. In the end, we briefly sketch possibilities for future work and try to encourage the reader to work on further improvements in the area of real-time depth imaging.

7.1 Summary

In the last chapters, we presented several approaches toward enhanced real-time depth sensing. We introduced the two key issues, namely the challenge of producing depth images in *real-time* and the verification that these measurements are *reliable*. We discussed which depth sensing methods and technologies are capable of providing depth images while meeting these demands. We concluded that ToF imaging is the most promising technology, as it enables the fast capture of reasonable large scenes. While, spatial resolution of these ToF cameras is meant to be increased in the next years, the technology suffers from a great many of systematic errors. We claim that a combination with other depth imaging systems will provide solutions that circumvent these systematic errors.

We propose to combine ToF cameras with stereo imaging in order to provide high resolution color information as well as depth information in a broader range of applications as a single sensor can provide. A proof of this concept has been described in Chapter 4, while an alternative and more practical solution has been demonstrated in Chapter 5. We combine a PMD and a stereo camera and control the disparity estimation with confidence data from the ToF sensor. We show that this is a crucial improvement for AR applications and can be seen as a first step toward the goal of applications like free 3D viewpoint television.

This enhances the range of applications toward a human-like machine vision, but suffers from a low dynamic range in the depth sensing. While a first, promising solution has been presented in Chapter 6, where a technique from computational photography has been successfully adapted to ToF imaging, the problem of the low dynamic range remains challenging. However, our solution is flexible enough to be adapted to the requirements in automotive engineering and can thus form the foundations for new driver assistance systems or even lead to autonomous vehicles.

7.2 Discussion

Impact This thesis contains several contributions to the scientific areas of CG and CV. We demonstrate the combination of depth sensing principles that might guide applied research toward devices that are utilized in unexplored terrains instead of the laboratory. Our approaches are meant to increase the generality of depth sensing and hence broaden the range of applications. One interesting application would be the embedding of such sensors in vehicles like the mars rover, where several sensors are attached together. There are no approaches to combine all these sensors in order to provide the rover with a deeper understanding of its surrounding. Our solutions show that combining depth sensors can improve the depth sensing capabilities, however it is not possible to test such a sensor completely for unknown environments.

Limitations The proposed solutions have only been tested in the laboratory. There is a big gap between standard testing environments and the real world. There might always occur unpredictable use cases that have not been tested and hence the functionality of approaches like ours can not be guaranteed. Besides this general limitation of devices tested in laboratory only, all our approaches introduce the benefit of reliability with a loss of speed. There is always a trade-off between precision and performance. While there will be future applications where our suggestions to improve the depth sensing do not work and the performance of the system might be reduced, our approaches do at least not decrease the reliability of the system.

7.3 Outlook

Combination One important aspect of this thesis is the proof that combining ToF and stereo imaging leads to significant improvements in terms of reliability of real-time depth imaging. We want to emphasize the importance of not focusing on a single sensing technology. Triangulation methods, for example, always suffer from the necessity of a baseline, while for ToF imaging it will always remain challenging to increase the signal to noise ratio while keeping the energy efforts as low as possible. We see great potential for saving energy also by combining methods and reducing computation costs. As an example the general combination of LIDAR and triangulation could lead to a setup where the reference signal of the

LIDAR system additionally forms a source of structured light that allows to triangulate. Then, the distance information can be extracted by two contrary methods having different noise characteristics, while the hardware costs are shared.

Integration Another outstanding aspect is the integration of methods from different fields into existing depth sensing technology. We demonstrated the possibility to enhance depth images by the help of an image fusion algorithm known from computational photography. In research on computational photography, further modifications of camera systems like coded aperture [66] show that there are plenty approaches to extract depth information from images. These findings could be further used to enhance the data captured by depth sensors like ToF cameras. In this thesis, a first approach from computational photography has been demonstrated and we encourage further research in this direction.

List of Figures

Bibliography

[1] A. Adams, M. Horowitz, S. H. Park, N. Gelfand, J. Baek, W. Matusik, M. Levoy, D. E. Jacobs, J. Dolson, M. Tico, K. Pulli, E.-V. Talvala, B. Ajdin, D. Vaquero, and H. P. a. Lensch. The frankencamera: an experimental platform for computational photography. *ACM Transactions on Graphics*, 29(4):1–12, July 2010.

[2] O. Arif, W. Daley, P. Vela, J. Teizer, and J. Stewart. Visual tracking and segmentation using time-of-flight sensor. In *Proceedings of International Conference on Image Processing (ICIP'10)*, pages 2241–2244, 2010.

[3] R. Bajcsy and L. Lieberman. Texture gradient as a depth cue. *Computer Graphics and Image Processing*, 5(1):52–67, March 1976.

[4] J. Bartl, R. Fíra, and M. Hain. Inspection of surface by the Moiré method. *Measurement Science Review*, 1(1):29–32, 2001.

[5] C. Beder, I. Schiller, and R. Koch. Real-time estimation of the camera path from a sequence of intrinsically calibrated PMD depth images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXVII:45–50, 2008.

[6] P. J. Besl. Active, optical range imaging sensors. *Machine Vision and Applications*, 1(2):127–152, June 1988.

[7] P. J. Besl and H. McKay. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[8] O. Bimber, T. Zeidler, A. Grundhöfer, G. Wetzstein, M. Möhring, S. Knödel, and U. Hahne. Interacting with augmented holograms. In *SPIE Conference on Practical Holography XIX: Materials and Applications*, 2005.

[9] Y. Boykov, O. Veksler, and R. Zabith. A variable window approach to early vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1283–1294, December 1998.

[10] D. E. Breen, R. T. Whitaker, E. Rose, and M. Tuceryan. Interactive occlusion and automatic object placement for augmented reality. *Computer Graphics Forum*, 15(3):11–22, 1996.

[11] 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. Technical report, CSEM, 2005.

[12] P. J. Burt and E. H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, COM-31,4:532–540, 1983.

[13] Y. Chen and G. Medioni. Object modeling by registration of multiple range images. In *Proceedings of IEEE International Conference on Robotics and Automation*, volume 3, pages 2724–2729, April 1991.

[14] Y. Cui, S. Schuon, D. Chan, S. Thrun, and C. Theobalt. 3D shape scanning with a time-of-flight camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*, pages 1173 –1180, June 2010.

[15] P. E. Debevec and J. Malik. Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH 2008 classes*, SIGGRAPH '08, pages 31:1–31:10, New York, NY, USA, 2008. ACM.

[16] O. Faugeras. *Three-Dimensional Computer Vision (Artificial Intelligence)*. The MIT Press, 1993.

[17] O. Faugeras, B. Hotz, H. Mathieu, T. Viéville, Z. Zhang, P. Fua, E. Théron, L. Moll, G. Berry, J. Vuillemin, P. Bertin, and C. Proy. Real time correlation based stereo: algorithm implementations and applications. Technical Report RR-2013, INRIA, 1993.

[18] R. Feris, R. Raskar, L. Chen, K. Tan, and M. Turk. Discontinuity preserving stereo with small baseline multi-flash illumination. In *IEEE International Conference in Computer Vision (ICCV'05)*, Beijing, China, 2005.

[19] J. Fischer, B. Huhle, and A. Schilling. Using time-of-flight range data for occlusion handling in augmented reality. In *Eurographics Symposium on Virtual Environments (EGVE)*, pages 109–116, 2007.

[20] S. Foix, G. Alenya, J. Andrade-Cetto, and C. Torras. Object modeling using a ToF camera under an uncertainty reduction approach. In *IEEE International Conference on Robotics and Automation*, pages 1306–1312. IEEE, May 2010.

[21] D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, 2002.

[22] M. Frank, M. Plaue, and F. A. Hamprecht. Denoising of continuous-wave time-of-flight depth images using confidence measures. *Optical Engineering*, 48(7):077003, 2009.

[23] M. Frank, M. Plaue, H. Rapp, U. Köthe, B. Jähne, and F. A. Hamprecht. Theoretical and experimental error analysis of continuous-wave time-of-flight range cameras. *Optical Engineering*, 48(1):013602, 2009.

[24] M. Franke. Color image segmentation based on an iterative graph cut algorithm using time-of-flight cameras. In R. Mester and M. Felsberg, editors, *Pattern Recognition*, volume 6835 of *Lecture Notes in Computer Science*, pages 462–467. Springer Berlin / Heidelberg, 2011.

[25] B. Freedman, A. Shpunt, M. Machline, and Y. Arieli. Depth mapping using projected patterns, October 2008. Patent: WO/2008/120217.

[26] J. García and Z. Zalevsky. Range mapping using speckle decorrelation, 2008. Patent: 7433024.

[27] J. García, Z. Zalevsky, P. García-Martínez, C. Ferreira, M. Teicher, and Y. Beiderman. Three-dimensional mapping and range measurement by means of projected speckle patterns. *Applied Optics*, 47(16):3032–3040, June 2008.

[28] S. Ghobadi. *Real Time Object Recognition and Tracking*. PhD thesis, Department of Electrical Engineering and Computer Science, Universität Siegen, 2010.

[29] J. Gibson. *The perception of the visual world.* Houghton Mifflin, 1950.

[30] B. E. Goldstein. *Wahrnehmungspsychologie - Der Grundkurs*. Heidelberg: Springer-Verlag Berlin, 2008.

[31] A. A. Goshtasby. Fusion of multi-exposure images. *Image and Vision Computing*, 23(6):611–618, June 2005.

[32] D. R. Griffin. *Listening in the dark: the acoustic orientation of bats and men*. New Haven : Yale Univ. Press, 1958.

[33] P. Grossmann. Depth from focus. *Pattern Recognition Letters*, 5:63–69, January 1987.

[34] S. Guðmundsson, H. Aanæs, and R. Larsen. Fusion of stereo vision and time-of-flight imaging for improved 3D estimation. In *International Workshop in conjunction with DAGM'07: Dynamic 3D Imaging.*, pages 164–172, Heidelberg, September 2007.

[35] R. Gvili, A. Kaplan, E. Ofek, and G. Yahav. Depth keying. In *Proceedings of SPIE*, volume 5006, pages 564–574. SPIE, 2003.

[36] U. Hahne and M. Alexa. Combining time-of-flight depth and stereo images without accurate extrinsic calibration. In *International Workshop in conjunction with DAGM'07: Dynamic 3D Imaging.*, pages 78–85, Heidelberg, September 2007.

[37] U. Hahne and M. Alexa. Combining time-of-flight depth and stereo images without accurate extrinsic calibration. *International Journal of Intelligent Systems Technologies and Applications*, 5(3/4):325–333, November 2008.

[38] U. Hahne and M. Alexa. Depth imaging by combining time-of-flight and on-demand stereo. In A. Kolb and R. Koch, editors, *Dynamic 3D Imaging*, volume 5742 of *Lecture Notes in Computer Science*, pages 70–83, Berlin, Heidelberg, 2009. Springer.

[39] U. Hahne and M. Alexa. Exposure fusion for time-of-flight imaging. *Computer Graphics Forum*, 30(7):1887–1894, September 2011.

[40] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[41] J. Heikkilae and O. Silven. A four-step camera calibration procedure with implicit image correction. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'97)*, pages 1106–1112, 1997.

[42] H. Hirschmüller. Improvements in real-time correlation-based stereo vision. In *Proceedings of the IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV'01)*, pages 141–148, Washington, DC, USA, 2001. IEEE Computer Society.

[43] H. Hirschmüller and D. Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:1582–1599, September 2009.

[44] A. Hornung and L. Kobbelt. Hierarchical volumetric multi-view stereo reconstruction of manifold surfaces based on dual graph embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 503–510, 2006.

[45] B. Huhle, T. Schairer, P. Jenke, and W. Straßer. Fusion of range and color images for denoising and resolution enhancement with a non-local filter. *Computer Vision and Image Understanding*, 114(12):1336–1345, December 2010.

[46] S. Hussmann, A. Hermanski, and T. Edeler. Real-time motion artifact suppression in ToF camera systems. *IEEE Transactions on Instrumentation and Measurement*, 60(5):1682–1690, May 2011.

[47] G. J. Iddan and G. Yahav. Three-dimensional imaging in the studio and elsewhere. In *Proceedings of SPIE*, volume 4298, pages 48–55. SPIE, 2001.

[48] A. Jones, M. Lang, G. Fyffe, X. Yu, J. Busch, I. McDowall, M. Bolas, and P. Debevec. Achieving eye contact in a one-to-many 3d video teleconferencing system. In *ACM SIGGRAPH 2009 papers*, SIGGRAPH '09, pages 64:1–64:8, New York, NY, USA, 2009. ACM.

[49] A. Jones, I. McDowall, H. Yamada, M. Bolas, and P. Debevec. Rendering for an interactive 360° light field display. In *ACM SIGGRAPH 2007 papers*, SIGGRAPH '07, New York, NY, USA, 2007. ACM.

[50] D. Jones and D. Lamb. Analyzing the visual echo: Passive 3-d imaging with a multiple aperture camera. Technical Report CIM-93-3, McGill University, Montreal, Canada, February 1993.

[51] T. Kahlmann, F. Remondino, and H. Ingensand. Calibration for increased accuracy of the range imaging camera swissranger. In H.-G. Maas and D. Schneider, editors, *Proceedings of the ISPRS Commission V Symposium 'Image Engineering and Vision Metrology'*, volume XXXVI, part 5, pages 136–141., Dresden, Germany, September 2006.

[52] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(9):920–932, 1994.

[53] M. Kanbara, T. Okuma, H. Takemura, and N. Yokoya. A stereoscopic video see-through augmented reality system based on real-time vision-based registration. *Proceedings of IEEE Virtual Reality (VR'00)*, pages 255–262, 2000.

[54] M. Keller and A. Kolb. Real-time simulation of time-of-flight sensors. *Simulation Modelling Practice and Theory*, 17(5):967–978, 2009.

[55] M. Keller, J. Orthmann, A. Kolb, and V. Peters. A simulation framework for time-of-flight sensors. In *International Symposium on Signals, Circuits and Systems (ISSCS'07)*, volume 1, pages 1–4, July 2007.

[56] W. N. Kellogg. *Porpoises and Sonar*. University of Chicago Press, 1963.

[57] K. Khoshelham. Accuracy analysis of kinect depth data. In D. Lichti and A. Habib, editors, *Inproceedings of International Society of Photogrammetry and Remote Sensing Workshop on Laser Scanning*, volume XXXVIII-5/W12, Calgary, Canada, August 2011.

[58] A. Kirmani, T. Hutchison, J. Davis, and R. Raskar. Looking around the corner using transient imaging. In *IEEE 12th International Conference on Computer Vision*, pages 159 –166, October 2009.

[59] H. Kjer and J. Wilm. *Evaluation of surface registration algorithms for PET motion correction*. Bachelor thesis, Technical University of Denmark, 2010.

[60] R. Koch, I. Schiller, B. Bartczak, F. Kellner, and K. Köser. Mixin3d: 3d mixed reality with ToF-Camera. In A. Kolb and R. Koch, editors, *Dynamic 3D Imaging*, volume 5742 of *Lecture Notes in Computer Science*, pages 70–83, Berlin, Heidelberg, 2009. Springer.

[61] A. Kolb, E. Barth, R. Koch, and R. Larsen. Time-of-flight sensors in computer graphics. In *Eurographics State of the Art Reports*, pages 119–134, 2009.

[62] V. Kolmogorov, R. Zabih, and S. J. Gortler. Generalized multi-camera scene reconstruction using graph cuts. In *Fourth International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*, 2003.

[63] T. Kreis. *Handbook of Holographic Interferometry: Optical and Digital Methods*. Wiley-VCH, 2005.

[64] K.-D. Kuhnert and M. Stommel. Fusion of stereo-camera and PMD-camera data for real-time suited precise 3D environment reconstruction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4780–4785, October 2006.

[65] R. Lange. *3D time-of-flight distance measurement with custom solid-state image sensors in CMOS/CCD-technology*. PhD thesis, University of Siegen, 2000.

[66] A. Levin, R. Fergus, F. Durand, and W. T. Freeman. Image and depth from a conventional camera with a coded aperture. In *ACM SIGGRAPH 2007 papers*, SIGGRAPH '07, New York, NY, USA, 2007. ACM.

[67] M. Levoy, K. Pulli, B. Curless, S. Rusinkiewicz, D. Koller, L. Pereira, M. Ginzton, S. Anderson, J. Davis, J. Ginsberg, J. Shade, and D. Fulk. The digital Michelangelo project: 3D scanning of large statues. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '00, pages 131–144, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co.

[68] Y. Li, S. Lin, H. Lu, S. B. Kang, and H.-Y. Shum. Multibaseline stereo in the presence of specular reflections. In *Proceedings of 16th International Conference on Pattern Recognition*, volume 3, pages 573 – 576, 2002.

[69] M. Lindner. *Calibration and Realtime Processing of Time-of-Flight Range Data*. PhD thesis, Department of Electrical Engineering and Computer Science, Universität Siegen, 2010.

[70] M. Lindner and A. Kolb. Lateral and depth calibration of PMD-distance sensors. *Sensors (Peterborough, NH)*, pages 524–533, 2006.

[71] M. Lindner and A. Kolb. Calibration of the intensity-related distance error of the PMD ToF-camera. In D. P. Casasent, E. L. Hall, and J. Röning, editors, *SPIE Conference Series: Intelligent Robots and Computer Vision XXV: Algorithms, Techniques, and Active Vision*, volume 6764. SPIE, September 2007.

[72] M. Lindner and A. Kolb. Compensation of motion artifacts for time-of-flight cameras. In A. Kolb and R. Koch, editors, *Dynamic 3D Imaging*, volume 5742 of *Lecture Notes in Computer Science*, pages 16–27. Springer, 2009.

[73] M. Lindner, A. Kolb, and K. Hartmann. Data-fusion of pmd-based distance-information and high-resolution rgb-images. In *International Symposium on Signals, Circuits and Systems (ISSCS'07)*, volume 1, pages 1 –4, July 2007.

[74] M. Lindner, A. Kolb, and T. Ringbeck. New insights into the calibration of ToF-sensors. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–5, Anchorage, AK, June 2008.

[75] M. Lindner, M. Lambers, and A. Kolb. Sub-pixel data fusion and edge-enhanced distance refinement for 2D/3D images. *International Journal of Intelligent Systems Technologies and Applications*, 5(3/4):344–354, November 2008.

[76] M. Lindner, I. Schiller, A. Kolb, and R. Koch. Time-of-flight sensor calibration for accurate range sensing. *Computer Vision and Image Understanding*, 114(12):1318–1328, December 2010.

[77] W. S. Malpica and A. C. Bovik. Range image quality assessment by structural similarity. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '09, pages 1149–1152, Washington, DC, USA, 2009. IEEE Computer Society.

[78] S. May, D. Droeschel, D. Holz, C. Wiesen, and S. Fuchs. 3D pose estimation and mapping with time-of-flight cameras. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Workshop on 3D Mapping*, Nice, France, October 2008.

[79] S. May, B. Werner, H. Surmann, and K. Pervolz. 3D time-of-flight cameras for mobile robotics. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 790–795, October 2006.

[80] T. Mertens, J. Kautz, and F. Van Reeth. Exposure fusion. In *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, pages 382–390. IEEE, October 2007.

[81] V. Mico, Z. Zalevsky, J. García, M. Teicher, Y. Beiderman, E. Valero, P. García-Martínez, and C. Ferreira. Three-dimensional mapping and ranging of objects using speckle pattern analysis. In P. Ferraro, A. Wax, and Z. Zalevsky, editors, *Coherent Light Microscopy*, volume 46 of *Springer Series in Surface Sciences*, pages 347–367. Springer Berlin Heidelberg, 2011.

[82] A. Mischler. 3D reconstruction from depth and stereo images for augmented reality applications. Diploma thesis, Technische Universität Berlin, December 2007.

[83] T. Möller, H. Kraft, J. Frey, M. Albrecht, and R. Lange. Robust 3D measurement with PMD sensors. In *Proceedings of the 1st Range Imaging Research Day at ETH*, pages 3 – 16, 2005.

[84] N. Naik, S. Zhao, A. Velten, R. Raskar, and K. Bala. Single view reflectance capture using multiplexed scattering and time-of-flight imaging. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, SA '11, pages 171:1–171:10, New York, NY, USA, 2011. ACM.

[85] S. Nayar, M. Watanabe, and M. Noguchi. Real-time focus range sensor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(12):1186–1198, 1996.

[86] C. Netramai, O. Melnychuk, J. Chanin, and H. Roth. Combining PMD and stereo camera for motion estimation of a mobile robot. In *The 17th IFAC World Congress*, July 2008.

[87] D. Nitzan, A. E. Brain, and R. O. Duda. The measurement and use of registered reflectance and range data in scene analysis. *Proceedings of the IEEE*, 65(2):206–220, 1977.

[88] T. Oggier, B. Büttgen, F. Lustenberger, G. Becker, B. Rüegg, and A. Hodac. Swissranger SR3000 and first experiences based on miniaturized 3D-ToF cameras. Technical report, CSEM, 2005.

[89] S. Paris, F. Sillion, and L. Quan. A surface reconstruction method using global graph cut optimization. *International Journal of Computer Vision*, 66(2):141–161, 2006.

[90] J. Penne, K. Höller, M. Stürmer, T. Schrauder, A. Schneider, R. Engelbrecht, H. Feußner, B. Schmauss, and J. Hornegger. Time-of-flight 3D endoscopy. In G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble, and C. Taylor, editors, *Medical Image Computing and Computer-Assisted Intervention MICCAI 2009*, volume 5761 of *Lecture Notes in Computer Science*, pages 467–474. Springer Berlin / Heidelberg, 2009.

[91] D. Piatti. *Time-of-Flight cameras: tests, calibration and multi-frame registration for automatic 3D object reconstruction*. PhD thesis, Politecnico di Torino, Italy, 2010.

[92] J. Radmer and J. Krüger. Making time-of-flight range imaging cameras applicable for human-machine-interaction by depth correction. In *3rd International Conference on Human System Interaction*, pages 393–398. IEEE, May 2010.

[93] J. Radmer, P. Moser Fusté, and J. Krüger. Distance calibration for ToF based range imaging sensors considering the amount of incident light. Technical report, Institute for Machine Tools and Factory Management, Technische Universität Berlin, 2008.

[94] H. Rapp. Experimental and theoretical investigation of correlating ToF-camera systems. Physics, Faculty for Physics and Astronomy, University of Heidelberg, Germany, September 2007.

[95] R. Raskar, A. Agrawal, and J. Tumblin. Coded exposure photography: motion deblurring using fluttered shutter. In *ACM SIGGRAPH 2006 Papers*, SIGGRAPH '06, pages 795–804, New York, NY, USA, 2006. ACM.

[96] E. Reinhard, G. Ward, S. Pattanaik, and P. Debevec. *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting (The Morgan Kaufmann Series in Computer Graphics)*. Morgan Kaufmann, 2005.

[97] R. Reulke. Combination of distance data with high resolution images. In H.-G. Maas and D. Schneider, editors, *Proceedings of the ISPRS Commission V Symposium 'Image Engineering and Vision Metrology'*, volume XXXVI, part 5, Dresden, Germany, September 2006.

[98] M. Reynolds, J. Dobos, L. Peel, T. Weyrich, and G. Brostow. Capturing time-of-flight data with confidence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*, pages 945–952, June 2011.

[99] J. B. T. M. Roerdink and A. Meijster. The watershed transform: definitions, algorithms and parallelization strategies. *Fundamenta Informatica*, 41:187–228, January 2000.

[100] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, 2002.

[101] I. Schiller. *Dynamic 3D Scene Analysis and Modeling with a Time-of-Flight Camera*. PhD thesis, University of Kiel, 2011.

[102] I. Schiller, C. Beder, and R. Koch. Calibration of a PMD-camera using a planar calibration pattern together with a multi-camera setup. *Proceedings of ISPRS Congress Beijing*, XXXVII:297–302, 2008.

[103] J. Schmidt, H. Niemann, and S. Vogt. Dense disparity maps in real-time with an application to augmented reality. In *Proceedings of 6th IEEE Workshop on Applications of Computer Vision (WACV'02)*, pages 225–230, 2002.

[104] M. O. Schmidt. *Spatiotemporal Analysis of Range Imagery*. PhD thesis, University of Heidelberg, Fakultät für Physik und Astronomie, 2008.

[105] K. Schreiber, J. D. Crawford, M. Fetter, and D. Tweed. The motor side of depth vision. *Nature*, 410(6830):819–822, April 2001.

[106] S. Schuon, C. Theobalt, J. Davis, and S. Thrun. High-quality scanning using time-of-flight depth superresolution. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–7. IEEE, June 2008.

[107] S. Schuon, C. Theobalt, J. Davis, and S. Thrun. LidarBoost: Depth superresolution for ToF 3D shape scanning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*, pages 343–350, June 2009.

[108] R. Schwarte, Z. Xu, H.-G. Heinol, J. Olk, and B. Buxbaum. New optical four-quadrant phase detector integrated into a photogate array for small and precise 3D cameras. In *Proceedings of SPIE*, volume 3023, pages 119–128. SPIE, 1997.

[109] A. Swadzba, B. Liu, J. Penne, O. Jesorsky, and R. Kompe. A comprehensive system for 3D modeling from range images acquired from a 3D ToF sensor. In G. Sagerer, editor, *Proceedings of International Conference on Computer Vision Systems*, pages 1–10, Bielefeld University, Bielefeld, Germany, March 2007. University Library of Bielefeld.

[110] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer, 2010.

[111] H. F. Talbot. LXXVI. facts relating to optical science, no. IV. *The London and Edinburgh Philosophical Magazine and Journal of Science*, 9(56):401–407, 1836.

[112] S. Tran and L. Davis. 3D surface reconstruction using graph cuts with surface constraints. In A. Leonardis, H. Bischof, and A. Pinz, editors, *Computer Vision (ECCV'06)*, volume Volume 3952 of *Lecture Notes in Computer Science*, pages 219–231. Springer Berlin / Heidelberg, 2006.

[113] B. van Loon. Radar 101: Celebrating 101 years of development. *Proceedings of the IEEE*, 93(4):844–846, April 2005.

[114] J. Ventura and T. Höllerer. Depth compositing for augmented reality. In *SIGGRAPH '08: ACM SIGGRAPH 2008 posters*, pages 1–1, New York, NY, USA, 2008. ACM.

[115] O. Vinyals, G. Friedland, and N. Mirghafori. Revisiting a basic function on current CPUs: a fast logarithm implementation with adjustable accuracy. Technical Report 510, International Computer Science Institute, Berkeley, California, Tech. Rep. TR-07-002, 2007.

[116] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, Apr. 2004.

[117] Wikipedia. Human echolocation — wikipedia, the free encyclopedia, 2011. [Online; accessed 12-December-2011].

[118] Wikipedia. Talbot effect — wikipedia, the free encyclopedia, 2011. [Online; accessed 30-January-2012].

[119] Wikipedia. Visual perception — wikipedia, the free encyclopedia, 2011. [Online; accessed 12-December-2011].

[120] Wikipedia. Calibration — wikipedia, the free encyclopedia, 2012. [Online; accessed 2-March-2012].

[121] Wikipedia. Differential signaling — wikipedia, the free encyclopedia, 2012. [Online; accessed 13-January-2012].

[122] Wikipedia. PID controller — wikipedia, the free encyclopedia, 2012. [Online; accessed 14-February-2012].

[123] Wikipedia. Time-of-flight camera — wikipedia, the free encyclopedia, 2012. [Online; accessed 5-February-2012].

[124] Z. Xu, R. Schwarte, H. Heinol, B. Buxbaum, and T. Ringbeck. Smart pixel: photonic mixer device (PMD). In *Proceeding of the International Conference on Mechatronics and Machine Vision*, pages 259–264, 1998.

[125] G. Yahav, G. Iddan, and D. Mandelboum. 3D imaging camera for gaming application. In *International Conference on Consumer Electronics (ICCE'07), Digest of Technical Papers*, pages 1–2, January 2007.

[126] T. Yu, N. Ahuja, and W.-C. Chen. SDG cut: 3D reconstruction of non-Lambertian objects using graph cuts on surface distance grid. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pages 2269–2276, Washington, DC, USA, 2006. IEEE Computer Society.

[127] Z. Zalevsky, A. Shpunt, A. Maizels, and J. Garcia. Method and system for object reconstruction, 2006. Patent: WO/2007/043036.

[128] Z. Zhang. A flexible new technique for camera calibration. Technical report, Microsoft Research, 1999.

[129] Z. Zhang. A flexible new technique for camera calibration. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(11):1330–1334, 2000.

[130] J. Zhu, L. Wang, R. Yang, and J. Davis. Fusion of time-of-flight depth and stereo for high accuracy depth maps. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'08)*, pages 1–8, June 2008.

[131] J. Zhu, L. Wang, R. Yang, J. Davis, and Z. Pan. Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1400–1414, July 2011.