Dissertation
submitted to the
Combined Faculties for the
Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

put forward by
Dipl.-Phys. Rahul Nair
born in Mumbai, India

Date of oral exam: February, 4th 2015

Analysis and Modeling of Passive Stereo and Time-of-Flight Imaging

Referees:
Prof. Dr. rer. nat. Bernd Jähne
Prof. Dr. rer. nat. Karl-Heinz Brenner

Zusammenfassung: Die vorliegende Arbeit befasst sich mit der Analyse und Modellierung von Effekten, die Fehler in passiver Stereoskopie und Laufzeitbildgebung verursachen. Die Hauptthemen sind in vier Kapiteln dargestellt: Ausgangspunkt ist eine Behandlung von Mischsystemen, die sich aus einer Laufzeitkamera und einem Stereosystem zusammensetzen. Einerseits zeige ich hierbei auf, wie häufig verwendete Fusionsansätze mit dem Messprozess der einzelnen Modalitäten zusammenhängen, andererseits präsentiere ich neue Techniken zur Datenfusion, welche die ermittelten Tiefenrekonstruktionen im Vergleich zu den einzelnen Sytemen verbessern können. Anschließend stelle ich ein System zur Erzeugung von großen Mengen von Referenzdaten für die quantitative Stereoevaluation vor, welches sich dadurch auszeichnet, dass neben Referenzgeometrien pro Pixel auch die Messunsicherheit der Referenzdaten erfasst wird. Die letzten beiden Teile umfassen schließlich Effekte, die in den einzelnen Systemen beobachtetet werden können: Laufzeitkameras können bekanntlich nur bis zu einem gewissen Abstand die Entfernung eindeutig bestimmen. Diesbezüglich zeige ich, dass durch die Änderung relevanter Designparameter des zugrunde liegenden Messsystems dieser Eindeutigkeitsbereich vergrößert werden kann. Zuletzt diskutiere ich, wie durch Modellierung eines begrenzten Lichttransports in der Szene es nicht nur möglich ist, systematische Fehler aufgrund von Reflektionen zu beheben, sondern auch einfache Materialparameter zu schätzen sowie die resultierenden Rekonstruktionen zu verbessern.

Abstract: This thesis is concerned with the analysis and modeling of effects which cause errors in passive stereo and Time-of-Flight imaging systems. The main topics are covered in four chapters: I commence with a treatment of a system combining Time-of-Flight imaging with passive stereo and show how commonly used fusion models relate to the measurements of the individual modalities. In addition, I present novel fusion techniques capable of improving the depth reconstruction over those obtained separately by either modality. Next, I present a pipeline and uncertainty analysis for the generation of large amounts of reference data for quantitative stereo evaluation. The resulting datasets not only contain reference geometry, but also per pixel measures of reference data uncertainty. The next two parts deal with individual effects observed: Time-of-Flight cameras suffer from range ambiguity if the scene extends beyond a certain distance. I show that it is possible to extend the valid range by changing design parameters of the underlying measurement system. Finally, I present methods that make it possible to amend model violation errors in stereo due to reflections. This is done by means of modeling a limited level of light transport and material properties in the scene.

For Mom and Pop

Acknowledgements

First and foremost, I thank my supervisor Prof. Dr. Bernd Jähne for making this work possible and for providing such a fruitful and stimulating lab environment at the Heidelberg Collaboratory for Image Processing (HCI) in general and in our group in particular. I am grateful for the support and scientific freedom I got to pursue my research.

I also thank Prof. Dr. Karl-Heinz Brenner for agreeing to be my second referee.

Dr. Daniel Kondermann has been a guiding spirit throughout these years and his enthusiasm (g.S.!) has helped me during several challenging periods. A lot of what I have learned in the past years, I owe to him and his motivating way: Be it the exploration of new topics, how to write papers or the more intricate aspects of joint collaborations.

My thanks are also due to the Ministry of Economy Baden-Württemberg and the companies Unexpected, Pixomondo, ScreenPlane, Bewegte Bilder and Tridelity for funding parts of my work under grant number 2-4225.16/380 in the research project 'Entwicklung von Systemen und Methoden zur effektiven Erstellung und Bearbeitung stereoskopischer Inhalte', which was jointly carried out with the Filmakademie Baden-Württemberg, Institute of Animation. To this end, I especially thank Simon Spielmann, Volker Helzle and Michael Bußler for the close cooperation, opening opportunities to test my methods and in general giving me a fresh perspective on the more practical aspects of vision systems that are intended to be used in the field. I would also like to acknowledge Johannes Peters for allowing us to test methods during the filming of 'How to Get a Girl in 60 Seconds' and Sebastian Cramer from ScreenPlane Gmbh on providing the opportunity to use their stereo rigs and sharing insights on the setup of camera systems.

I am grateful to Dr. Christoph Garbe and Dr. Frank Lenzen for inviting me to join the 'Algorithms for Low Cost Depth Imag-

ing' joint project between the HCI and the Intel Visual Computing Institute in Saarbrücken as an associate member. Many a Wednesday meeting with interesting discussions will be remembered fondly. I am additionally indebted to Frank for introducing me to variational methods and for having the patience to discuss and explain things the many times I sought him out for help. I also thank Dr. Florian Becker, who I often approached for advice on various mathematical topics.

During the course of my work, I had the opportunity to work as a research intern at Microsoft Research Cambridge, UK. What started out as a limited three month project ended up as a long lasting collaboration, not least due to the amazing mentoring by Prof. Dr. Carsten Rother and Dr. Andrew Fitzgibbon. I thank them for reflecting on reflections with me, motivating me when I was on the verge of giving up this topic and also for the occasional but well placed 'this is not the right way to do it'.

My gratitude also goes to the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences (HGS MathComp) for funding many of my travel expenses and providing such a great platform that fosters the exchange of knowledge between doctoral students of many fields. I greatly enjoyed and profited from the encounters I had there. To this end, I would especially like to thank Dr. Michael Winckler and the rest of the organizing team.

All members of my lab have been a source of great camaraderie both at professional and social level, especially Dr. Stephan-The-Swiss-Army-Knife Meister, Jens-Malte Gottfried, Dr. Henrik Schäfer, Katrin Honauer, Burkhard Güssefeld, Moritz Becker, Gerald Mwangi, Michael Baron, Dr. Sven Wanner and Max Diebold. My dissertation would not have been possible without the friendly support and in-depth discussions on- and off-topic on- and off-hours.

Many thanks also go to Karin Kruljac, Barbara Werner, Evelyn Wilhelm and Dr. Christian Schmidt for doing such an amazing job of keeping all of our heads free of the administrative fuss.

I thank the people who proof-read parts or my whole thesis: Cornelia Herkommer, Urmila Nair, Matthias Henrich, Daniel Kondermann, Stephan Meister, Katrin Honauer, Burkhard Güssefeld, Frank Lenzen, Florian Becker, Mirko Schmidt, Carsten Rother, Andrew Fitzgibbon and Bernd Jähne.

Furthermore, I would like to thank the people who influenced

my vocational choices: I am grateful to the teachers I had in school who provided a secure foundation that I could build upon during my later studies: Ms. Wirth-Brunner, Ms. Brandt, Mr. Hoffmann, Mr. Knauf, Mr. Taufertshöfer and Ms. Grimm. Many thanks are due to Prof. Dr. Fred Hamprecht for giving lectures that made me enthusiastic about imaging physics, image processing and machine learning. I also thank him for those (initially) dreaded Tuesdays, where in retrospect I was taught how to efficiently read papers. My thanks also go to Dr. Ullrich Köthe, who deepened my interest in machine learning methods and taught me C++ at a time I had the arrogance of believing I knew everything there was to it. Finally, I thank Dr. Mirko Schmidt, who steered me towards depth imaging when I was pondering which topic to pursue in my doctoral studies.

On a more personal level, I thank my parents Dr. Urmila Nair and Dr. Jagadeesan Nair for encouraging my curiosity and teaching me what it means to be a scientist - even off hours. I am grateful for all my friends and family, who added that bit of fairy dust to the equation, making even the dimmest hours seem bright. Last but not least, I would like to thank my dear Conny for her patience, incredible proof-reading skills and providing 24/7 moral support in times of need.

Contents

1 Introduction

1.1 General Introduction

When do early vision techniques fail? How can we quantify these failures and finally, what can we do to correct them? These are the three driving questions behind the work presented in this thesis. But before these questions can be answered, it needs to be clarified what early vision is in the first place.

In 1985, Poggio et al. reviewed the developing field of computational vision and defined early vision as

> *the set of visual modules that aim to extract the physical properties of the surfaces around the viewer, that is distance, surface orientation and material properties [...].[146]*

This is in contrast to high-level vision, which is concerned with image semantics, such as object detection [62, 182, 50], classification [116, 22] or segmentation [167, 138, 52]. Today, the application domains (cf. Figure 1.1) for vision techniques are quite diverse and include human-computer interaction [176], automotive systems [7], industrial inspection [80], robotics [44], remote sensing [8], augmented reality [135] and visual effects [171]. The quote above originally referred to methods that mimic human vision such as passive stereo [93], structure from motion [100] or

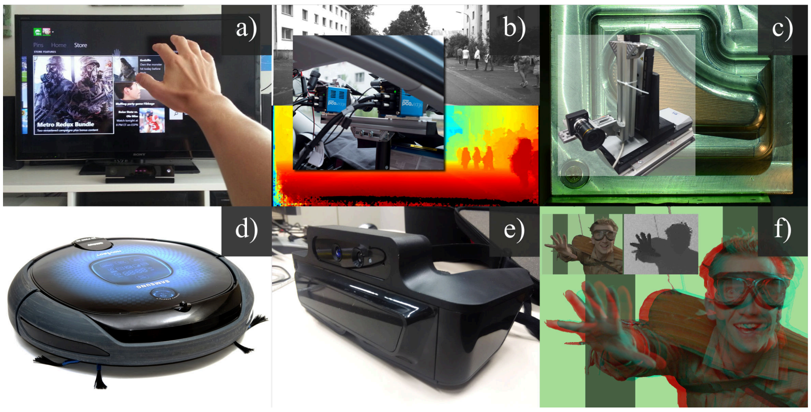


Figure 1.1: Examples for Applications That Make Use of Early Vision. From top left to bottom right: **(a)** ToF-based second generation Kinect used as a controller. **(b)** Stereo setup used in a car for pedestrian localization. **(c)** Surface inspection of industrial parts using a camera mounted on a gantry (Courtesy Sven Wanner/Max Diebold) **(d)** Self localizing vacuum cleaning robot. **(e)** Augmented reality with glasses equipped with a ToF sensor and **(f)** Input from a ToF and a single standard camera used to generate the second view for stereoscopic presentation.

shape from shading [148]. Nowadays, the field also encompasses other principles including structured light [158] and Time-of-Flight (ToF) imaging [177]. All these methods primarily focus on geometry extraction and make assumptions about lighting as well as material properties to make any reconstruction viable. Indeed, most efforts to recover material properties using imaging techniques have been driven by the computer graphics community. Here, methods were created to extract material properties by special instrumentation [129] or by using inverse rendering techniques [142]. Again, assumptions have to be made - this time about the geometry or lighting.

The work presented in this thesis is concerned with the analysis and modeling of early vision problems present in ToF imaging and passive stereo. The emphasis on 'early vision problems' is made to distinguish between the topics in this thesis from related fields of sensor or camera characterization [48, 163], which are beyond the scope of this work. The topics presented are however, in parts, interdisciplinary in nature as I often borrow concepts or methods from the related fields of computer graphics (Chapter 6), experimental design (Chapter 5) and photogrammetry (Chapter 4).

Despite their success, passive stereo and ToF imaging remain far from perfect as there still remain many situations where the methods fail. Therefore, a considerable amount of effort is put into amending the acquired data, either by manual intervention, by filtering or by adding high-level information into the reconstruction process (by means of regularization). The emphasis in this thesis is not on such robustification strategies, but rather on accounting for errors by modeling the effects that cause them.

The issues encountered fall roughly into three well-known categories and will be briefly illustrated using an example of estimating parabola parameters from measured positions, e.g. to estimate the parameters of a ballistic trajectory.

Ambiguities Consider the problem of fitting a parabola to two measurements (cf. Figure 1.2). This problem is underdetermined such that there are an infinite number of possible solutions. It is therefore not possible to estimate the right parameters without inserting any prior knowledge, e.g. starting point or angles of the trajectory. In this example, the ambiguity can be resolved if a third independent measurement is added. In the work presented, such ambiguities are encountered in stereo matching on non-textured surfaces (cf. Figure 1.5) and in ToF imaging as the range ambiguity problem (cf. Figure 1.7).

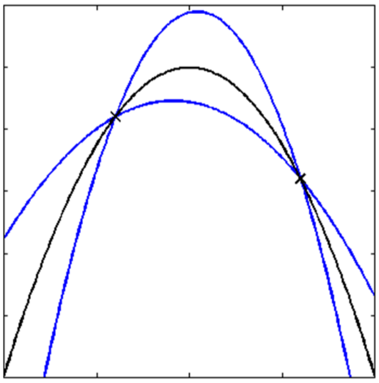


Figure 1.2: Illustration of Infinite Solutions Due to Data Ambiguity. The black curve represents the true trajectory while the blue curves correspond to two possible solutions given only two measurements.

Statistical Errors Next, there are errors that arise due to measurement uncertainty (cf. Figure 1.3). Any quantity that is measured is subject to such errors due to sensor noise or limited resolution. Since they are of statistical nature, no two measurements of the same quantity will yield precisely the same results. These measurement errors are propagated to the parameter fit such that the obtained curve parameters also deviate from the true ones. Since it is impossible to fully eliminate the statistical errors of measurements, it is of utmost importance to be able to assess their influence on the resulting parameter estimates. While also considered in other parts of my thesis, this aspect is most prominently featured in the uncertainty analysis of reference data generation (cf. Figure 1.6), where the influence of measurement errors on the quality of reference data is investigated.

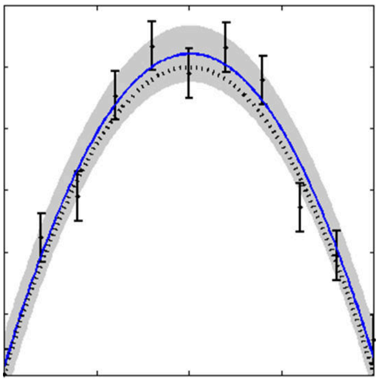


Figure 1.3: Illustration of Measurement Uncertainty. Noise in measurements causes a deviation between true (black dotted line) and fitted trajectory (blue). By propagation of uncertainty, it is possible to estimate the confidence (gray area) of the fit.

Systematic Errors due to Model Violations The third class of errors is related to model violations. So far, in our example, air drag has been left unconsidered while modeling the trajectory of the projectile. While it may be barely noticeable at low projectile velocities, air drag can significantly alter the trajectory from the parabola shape at higher velocities (cf. Figure 1.4). If not taken into account, a considerable systematic deviation can be observed between measured and computed trajectory. Comparable situations also occur in computer vision. Since most reconstruction formulas assume that all visible surfaces behave like Lambertian reflectors, i.e. appear the same irrespective of viewing angle, such model violation errors can occur whenever

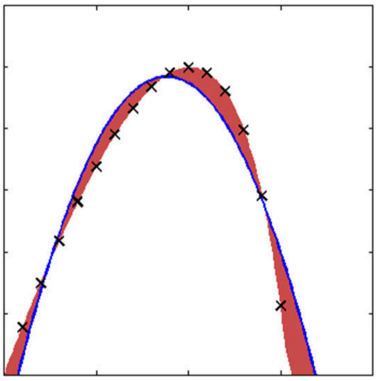


Figure 1.4: Illustration of Systematic Errors(red) Due to Model Violations. These are introduced if a quadratic model function (blue) is fitted to measurements of a ballistic projectile that is additionally subject to air drag (black markers).

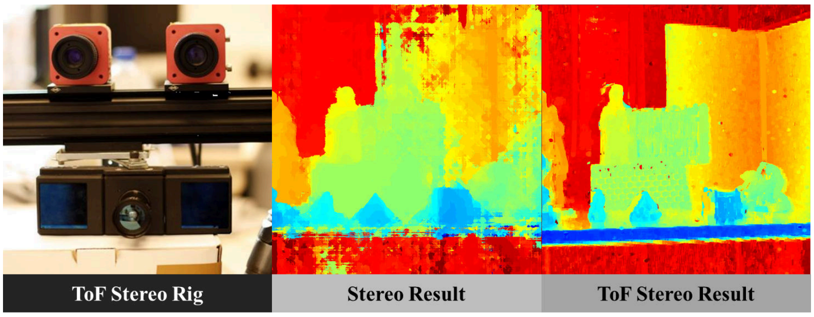


Figure 1.5: ToF-Stereo Fusion. **Left**: ToF-Stereo rig. **Middle**: Scene reconstruction using stereo only with errors due to lacking texture. **Right**: ToF-Stereo reconstruction.

reflective surfaces are present as they have a viewpoint dependent appearance (cf. Figure 1.8).

1.2 Topics

The four topics of this thesis *Time-of-Flight Stereo Fusion*, *Reference Data with Uncertainty*, *Time-of-Flight Range Extension* and *Reflections on Stereo* directly relate to the error sources discussed above and are summarized here:

ToF-Stereo Fusion The motivation for the first chapter of this thesis emerged from the observation that both ToF imaging and depth from stereo have issues inherent to their respective measurement principles, but the situations where these errors occur are often different. Thus, I investigated whether joint measurements using both modalities could allow for more robust depth reconstructions and indeed, I can present a ToF-Stereo fusion system that displays this behavior (cf. Figure 1.5). Additionally, I show how the method presented as well as the majority of related work can be derived from the least squares formulation of each individual method by a series of approximations and modifications. The evaluation of the methods presented was undertaken using measured ground truth data - that is by comparing the algorithm output with reference data generated by other means. To this end, I present one of the first publicly available measured evaluation datasets for ToF-Stereo fusion with ground truth.

Reference Data with Uncertainty One observation made during the work on sensor fusion is that errors in the pose estimates (i.e. the relative translation and rotation) between two measurement systems lead to alignment errors that are challenging

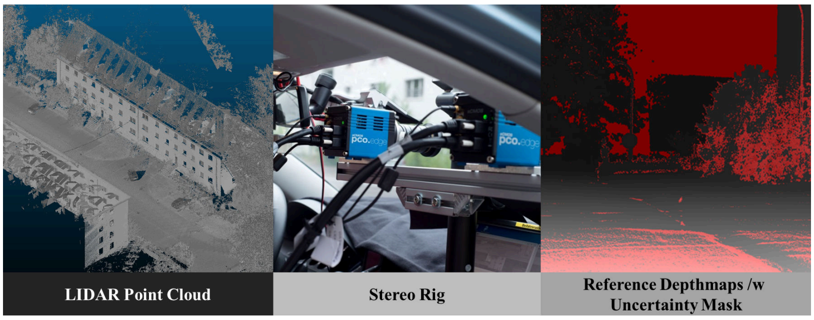


Figure 1.6: Reference Data with Uncertainty. **Left**: LIDAR point cloud used to generate stereo reference data. **Middle**: Stereo rig used for image acquisition. **Right**: Depth maps with a mask excluding pixels beyond a certain uncertainty.

to handle. The quantification of these errors with application to reference data generation for stereo evaluation is the goal of the second topic (cf. Figure 1.6). Here the output of a stereo algorithm is compared against range measurements made by a LIDAR system for evaluation of algorithm performance. Inevitable errors in the pose estimate between the stereo and the LIDAR coordinate frames together with the other present measurement errors lead to an uncertainty in the reference data. This uncertainty has to be accounted for when any kind of quantitative performance analysis is the goal. Yet to date, little work has been done on correctly extracting uncertainties for this kind of reference data generation. Hence, I present a pipeline that enables the production of reference data with per pixel uncertainty estimates. Furthermore, I show how performance analysis can benefit from such uncertainties.

ToF Range Extension Phase based ToF cameras recover depth by estimating the offset, amplitude and phase of a cosine function with a fixed frequency that is sampled at 4 (or more) locations. The estimated phase can then be converted into a distance using a linear transform. It is therefore only determined up to multiples of 2π causing cyclic errors if the scene extends beyond the distance corresponding to a phase of 2π (cf. Figure 1.7). In Chapter 5, I revisit the least squares formulation

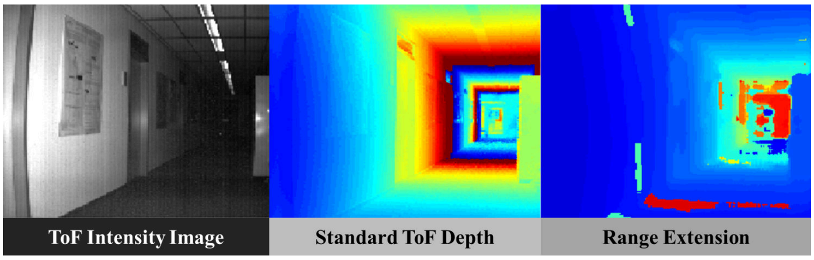


Figure 1.7: ToF Range Extension. **Left**: ToF intensity image. **Middle**: Range images using standard reconstruction. Note the cyclic depth errors caused by range ambiguity. **Right**: Reconstructions without ambiguity using the proposed method.

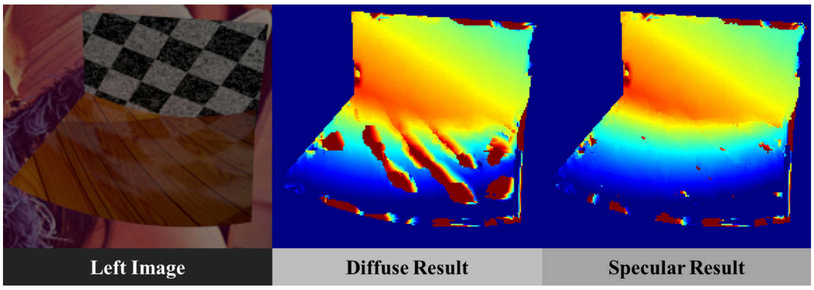


Figure 1.8: Reflections on Stereo. **Left**: Left image of a stereo pair of a curved surface with reflections. **Middle**: Depth reconstruction using standard stereo matching. **Right**: Depth reconstruction accounting for specular surfaces.

for ToF parameter estimation and show that these ambiguities naturally resolve if the modulation frequency between individual measurements per pixel is varied - amounting to changing the design variables in the least squares problem. Unlike existing work, the method presented in this thesis neither increases uncertainty of the estimated parameters nor does it require additional measurements. Also, it does not depend on any kind of regularization.

Reflections on Stereo Reflections on specular surfaces can cause large errors in both ToF and Stereo that current data fusion techniques cannot cope with. The errors caused by such surfaces belong to the category of model-violation errors since all traditional methods for both ToF and Stereo reconstruction assume that all visible surfaces are Lambertian. For a mirroring surface this is obviously not the case. Handling these errors is especially challenging as they are a) scene dependent and b) caused by highly nonlocal effects. Chapter 6 focuses on resolving the model violation errors for stereo matching in a theoretically sound manner. Unlike most existing work which employs regularization or robust data terms to suppress such errors, I derive two least squares models from first principles that generalize diffuse world stereo and explicitly take reflections into account. These models are parameterized by depth, orientation and material properties, resulting in a total of up to 5 parameters per pixel that have to be estimated. Additionally, large nonlocal interactions between viewed and reflected surfaces have to be taken into account. These two properties make model inference appear prohibitive at first, but I present evidence that it is actually possible. Finally, results indicate that the information gained by reflections actually leads to better reconstructions compared to the case where no reflections were present.

1.3 Road Map

Following the four main areas of contribution introduced above, this thesis is organized in four main parts enframed by a background and a concluding chapter. In the next chapter **(Chapter 2)**, I will review the theory and techniques that form the basis for the work presented later. To stay in scope, I will limit myself to topics essential towards understanding the presented matter and refer to standard literature for further reading. Any additional background information specific to one of the main chapters will be presented in the respective areas. Once set up with the required instruments, **Chapters 3 through 6** will present the main body of work. While each chapter can be read stand-alone together with the background chapter, I deemed this ordering best for understanding how the topics are associated with each other. Each chapter starts off with introductory sections explaining motivation, key contributions and discussion of related work before models, methods and results are presented. The chapters then conclude with a discussion of future work and a summary of the insights gained. In some cases, the future work section may also contain results of preliminary experiments that evidence the utility of the proposed ideas. In **Chapter 3**, I will first present work concerned with ToF-Stereo that serves as the motivation for the subsequent chapters. **Chapter 4** deals with the uncertainty estimation and propagation for the alignment of range and stereo data with application to reference data generation. The next two chapters depart from the dual-modality setup and deal with ToF and stereo systems individually: In **Chapter 5**, I discuss a design based approach towards extending the measurement range of a ToF camera whereas **Chapter 6** is concerned with understanding and handling errors on reflective surfaces by material estimation. In the final chapter **(Chapter 7)**, I will review the main results presented and discuss the overall lessons learned during the course of my work.

2 Background

As early vision is essentially about inverting the image formation process, I commence by presenting the key aspects that govern image formation in Section 2.1. Section 2.2 then reviews the two depth imaging techniques that this thesis is concerned with: passive stereo and Time-of-Flight (ToF) imaging. Finally, Section 2.3 concludes this chapter with a treatment of parameter estimation techniques. These methods are required ubiquitously for recovering scene parameters from observed images and form the basis for the methods presented.

The level of detail of the topics discussed here was chosen with the aim of putting the later chapters into context. Moreover, I refer to standard literature throughout the chapter for a more complete treatment of each topic.

2.1 Image Formation

Following [85], there are three different aspects that contribute to the formation of a digital image. Figure 2.1 illustrates these parts. A digital image is in essence a projection of the 3D world onto a 2D surface, namely the camera sensor. The positions $(p_x, p_y) =: \mathbf{x} \in \Omega \subset \mathbb{R}^2$ on the sensor surface plane that a world point $(x, y, z) =: \mathbf{X} \in \mathbb{R}^3$ projects onto, are defined by the **geometric and optical** properties of the camera system. The amount of light that $\mathbf{x}$ receives from $\mathbf{X}$ is governed by scene

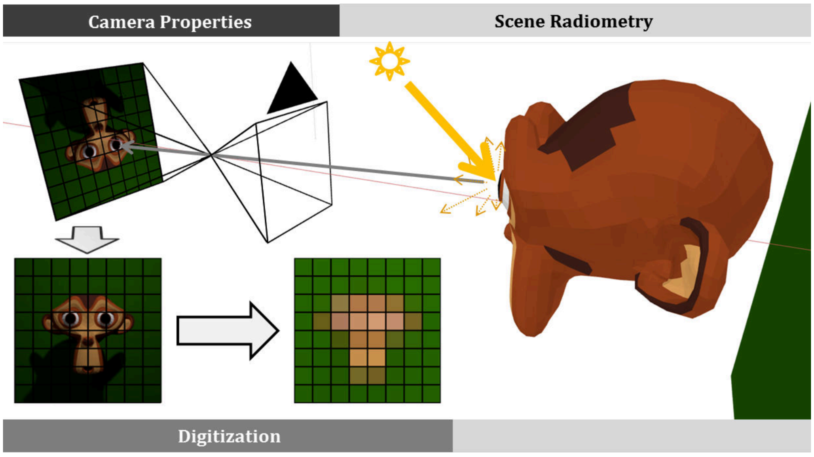


Figure 2.1: Image Formation in a Nutshell. Scene radiometry governs how light is transported in the scene and defines the appearance of the monkey's head (Suzanne [18]). The properties of the camera system define how the light emitted from the scene is mapped onto the sensor plane of the camera leading to an intensity distribution on the sensor plane. This intensity distribution is sampled on a regular grid of pixels and the signal is further quantized to obtain a digital image.

radiometry, which explains the light transport in the scene. It should be noted that the distinction between camera properties and scene radiometry is mostly conceptual. The same light transport laws that govern the amount of light directed into the camera also explain the image created on the camera sensor. This is illustrated in Figure 2.2, where a camera image is simulated by including the camera as part of the scene description. However, for image processing purposes it is more useful to model cameras separately from the rest of the scene and define the camera using few system parameters.

Once light is transported from the scene onto the camera, an image is formed on the sensor as a continuous distribution of incident radiance. This intensity distribution is then sampled on the (mostly rectangular) grid of light sensitive pixel sensors and finally converted into a **digital** signal consisting of fixed size (e.g. 8, 12 or 16 bit) integers.

For color images, three intensity distributions corresponding to red, green and blue (RGB) wavelengths are sampled separately. Depending on design, each pixel location may either measure all three colors simultaneously [82] or only measures one of the three colors, which occurs more frequently. For the latter case, a dense RGB image is recovered by an interpolation process called de-bayering or de-mosaicing[99].

In the following, each of the three aspects: camera properties, scene radiometry and image digitization will be discussed in further detail with a focus on the first two parts as they are most relevant towards the work presented.

2.1.1 Camera Models

Let $\mathbf{x} \in \Omega$ be the position on the image sensor plane Ω and $\mathbf{X} \in \mathbb{R}^3$ be a point in 3D space. The camera model describes the mapping between $\mathbf{X}$ and $\mathbf{x}$. Most generally, the projection of $\mathbf{X}$ onto Ω is a distribution

$$\mathrm{PSF}^{\mathbf{X}} : \Omega \to \mathbb{R}, \tag{2.1}$$

with

$$\int_{\Omega} \mathbf{dx}\, \mathrm{PSF}^{\mathbf{X}}(\mathbf{x}) = 1. \tag{2.2}$$

This distribution is called the point spread function (PSF) of the optical system and it describes how a point in space is imaged on the sensor plane. Now, let

$$\pi_d : \mathbb{R}^3 \to L^p(\mathbb{R}^2), \pi_d(\mathbf{X}) = \mathrm{PSF}^{\mathbf{X}}, \tag{2.3}$$

describe the projection of $\mathbf{X}$ onto Ω and let

$$\pi : \mathbb{R}^3 \to \mathbb{R}^2, \pi(\mathbf{X}) = \operatorname*{argmax}_{\mathbf{a}} \mathrm{PSF}^{\mathbf{X}}(\mathbf{a}), \tag{2.4}$$

map the 3D point to the mode of the distribution. π may depend on some additional camera parameters $\theta \in \mathbb{R}^N$. This dependency is made explicit as $\pi(\theta, \mathbf{X})$ wherever it is required but otherwise omitted for legibility. Note that π is not bijective as depth information is lost due to the projection. The actual form of the PSF additionally depends on the internal camera geometry of the optical system and may also be wavelength dependent (e.g chromatic aberrations).

Different light and camera models with varying complexity exist that approximate the true PSF of an optical system. The most frequently used model in early vision is that of a pinhole

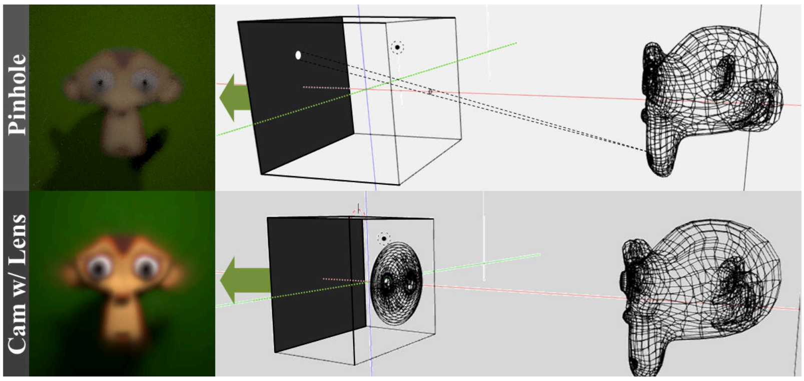


Figure 2.2: Simulating Cameras as Part of the Scene. In these examples, the cameras (top: pinhole with finite aperture, bottom: thick lens) were modeled as any other part of the scene. The images on the left are the intensity distributions that appear on the dark gray surfaces after simulating the light transport in the scene. The images are blurred due the spread of the projection of a single point in space given by the PSF (white circle in top row).

camera combined with ray optics, which can describe the basic projective properties of a large variety of cameras. The PSF of an ideal pinhole camera is a δ distribution such that π contains all information of this optical system. More complex models are based on this mapping to explain depth-of-field effects, distortions and other lens aberrations [1]. Finally, if the wave nature of light is taken into account, diffraction effects of the optical system can also be modeled.

The choice of model depends on the application and the required expressiveness. For example, Depth-from-defocus techniques [161] that estimate depth from the level of blurriness will require a thin lens-model that explains depth-of-field. Similarly, blind deconvolution [6] techniques that aim to de-blur images may need to estimate the Airy-disk PSF that can be modeled by wave optics. For the purposes of this work, the pinhole camera model with radial distortions suffices as it can model all relevant aspects of the problems that are considered. Further information on PSFs and their derivation can be found in [85] and [64](where it is called ‘impulse response’).

Pinhole Camera

For a pinhole, the projective mapping π is defined as

$$\pi(\mathbf{X}) = \pi(x, y, z) = \left(f_x \frac{x}{z} + c_x, f_y \frac{y}{z} + c_y \right). \tag{2.5}$$

Here it is assumed that the camera is placed at the origin of the world coordinate frame and views in positive z direction. A 2D example is illustrated in Figure 2.3. (f_x, f_y) and (c_x, c_y) are camera model parameters and are called the intrinsic parameters of the camera. (f_x, f_y) are the horizontal and vertical focal lengths of the camera and define the image magnification in those directions. While in this thesis it is $f_x = f_y =: f$, the two parameters may be different in the general case. Reasons for this are special lens geometries or if the pixel pitch, i.e. the distance between pixels on the camera, is not the same in both directions. The principal point (c_x, c_y) defines the image coordinates that the optical axis (i.e. all 3D points with $x = y = 0$) maps onto.

All the points in space that would be mapped onto the same

[1] In these models, $\pi_d(X)$ can often be expressed as a single distribution centered around the pinhole projection location.

image coordinates lie on a ray that intersects the pinhole.

$$\pi\left((0,0,0)+\lambda(r_x,r_y,r_z)\right)=\left(f\frac{r_x}{r_z}+c_x, f\frac{r_y}{r_z}+c_y\right). \tag{2.6}$$

The points actually observed depend on transmissivity and reflectivity of each of these points. Furthermore, straight lines given by $(s_x,s_y,s_z)+\lambda(r_x,r_y,r_z)=\mathbf{s}+\lambda\mathbf{r}$ that do not intersect the camera center are mapped onto straight lines in the image:

$$\pi(\mathbf{s}+\lambda\mathbf{r})=\pi(\mathbf{s})+\gamma\left(\pi(\mathbf{s}+\mathbf{r})-\pi(\mathbf{s})\right), \tag{2.7}$$

with

$$\gamma=\frac{\lambda(r_z+z)}{z+\lambda r_z}. \tag{2.8}$$

Finally, projections of parallel lines all intersect in a single image point called vanishing point. For fixed $\mathbf{r}$ and arbitrary $\mathbf{s}$ this can be seen by computing

$$\lim_{\lambda\to\infty}\pi(\mathbf{s}+\lambda\mathbf{r})=(f\frac{r_x}{r_z}+c_x, f\frac{r_y}{r_z}+c_x)=\pi(\mathbf{r}). \tag{2.9}$$

This position corresponds to the projection of the parallel line that intersects the camera center.

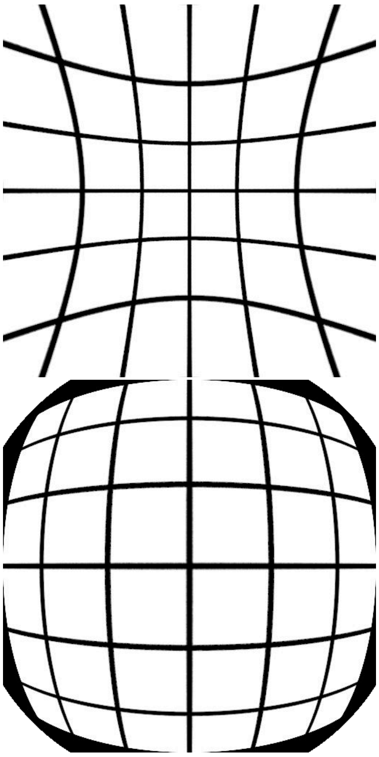


Figure 2.4: Barrel and pincushion distortions cause straight 3D lines to no longer be mapped onto straight lines in the image. Images are usually corrected in regard to these distortions prior to any further processing.

Radial Distortions

Real cameras have a lens instead of a pinhole. For the purposes of this work, it is sufficient to approximate the camera by the pinhole model if the sensor is in focus. That is, if the sensor is placed at the distance from the optical center, at which a sharp image, designated the focal length f of the lens, is formed. Also the aperture, i.e. the opening of the camera, has to be sufficiently

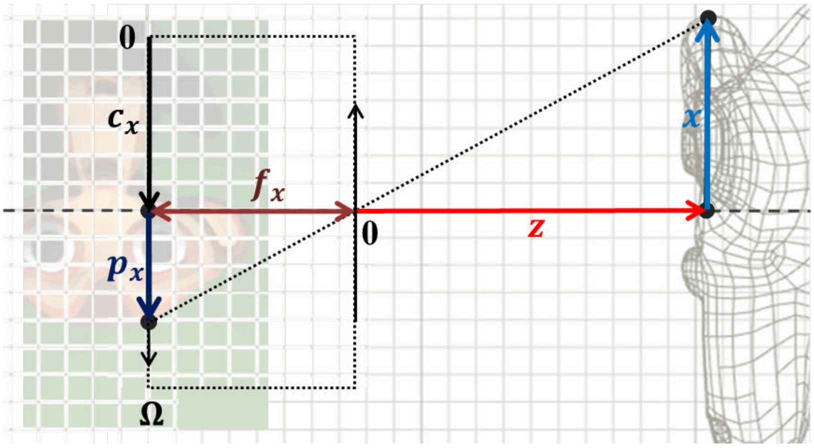


Figure 2.3: Ideal Pinhole Camera in Two Dimensions. The world coordinate origin is set in the center of the camera, which contains a infinitesimally small pinhole. For each point in space (x,y) this pinhole only allows light along a single ray direction to fall onto the sensor position p_x. The mapping between (x,y) and p_x depends on f_x and is defined by the intercept theorem. Finally, note that the origin of the image plane need not coincide with the optical axis but has the coordinate c_x .

small. Finally, barrel or pincushion distortions (cf. Figures 2.4) have to be taken into account. These distortions occur due to the projection of a real optical system deviating from the pinhole model and are present in the majority of lens/camera systems that were used in this work. As camera lenses often possess a radial symmetry, the distortions most observable have a radial symmetry as well. If $\mathbf{x}$ is the image coordinate the pinhole camera projects onto, then the real projected point $\mathbf{x}'$ is displaced in radial direction according to a function of the distance between $\mathbf{x}$ and (c_x, c_y). Functions commonly used to model these distortions are symmetric polynomials or rational functions. The parameters of these functions also belong to the intrinsic camera parameters θ. In practice and to simplify computations, the distortions parameters are estimated along with focal length and principal point during calibration. Then the image is undistorted [85] in a pre-processing step to obtain images as viewed through a pinhole camera.

Camera Extrinsics

The origin of the world coordinate system was assumed to coincide with the camera center till now. In reality, the measurement setup used often suggests a world coordinate system where neither origin nor orientation coincide. As an example, this is the case in a multi-camera setup, where the coordinate system of one camera is used as the world coordinate system for all others. The two different frames of references can be transformed into one another by means of a rotation of the coordinate axes around the origin and a translation of the coordinate origin (cf. Figure 2.5). A vector $\mathbf{X}_w$ described in terms of the world coordinate system w can then be expressed in terms of the camera system c as

$$\mathbf{X}_c = \mathbf{R}\mathbf{X}_w + \mathbf{s}, \tag{2.10}$$

given rotation matrix $\mathbf{R}$ and shift $\mathbf{s}$. The tuple $\mathbf{t} = (\mathbf{R}, \mathbf{s})$ is called the extrinsic parameters or the pose of the camera relative to the world coordinate system. using Eq. (2.5), I define

$$\pi(\theta, \mathbf{X}^w, \mathbf{t}) = \pi(\theta, \mathbf{R}\mathbf{X}_w + \mathbf{s}), \tag{2.11}$$

as the extended mapping that takes the pose of the camera into account. Again, the dependency may be omitted or added as superscript for reasons of clarity.

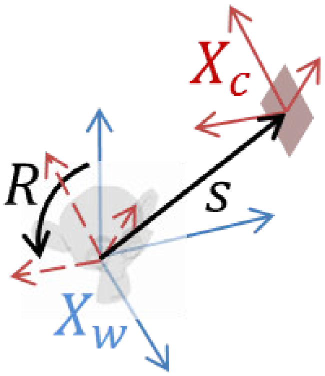


Figure 2.5: Illustration of Camera Extrinsics. A world coordinate system is transformed into the local camera coordinate system by means of a rotation $\mathbf{R}$ and a shift $\mathbf{s}$.

Rotation matrices are not the only representation of rotations available. For parameter estimation purposes, a compact representation of rotations is more useful as it better expresses the structure of the lower dimensional manifold of rotations. In the work presented, I use a representation of rotations in terms of a rotation vector $\mathbf{r} \in \mathbb{R}^3$, where the direction of $\mathbf{r}$ defines an axis about which a point is to be rotated. In addition, the length of the vector $||\mathbf{r}||$ corresponds to the angle which the point is rotated around the axis (cf. Figure 2.6) Any rotation matrix can be converted into a rotation vector and vice versa using Rodrigues' formula [121].

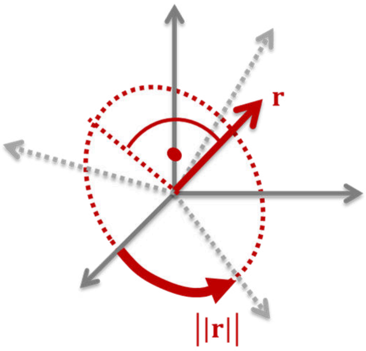


Figure 2.6: Angle Axis Representation of a Rotation. The direction of the rotation vector r defines an axis around which coordinates are rotated. The magnitude of r is the angle of rotation.

Calibration

Depth imaging is used to restore the 3D coordinates from (multiple) 2D projections. For this purpose it is essential to recover the intrinsic and extrinsic camera parameters. It is usually done by imaging special targets where the 3D positions $\mathbf{X}_i$ of some landmark points are known in the object coordinate system. Similarly, the projected positions $\mathbf{x}_i$ can be easily extracted from the image. The unknown intrinsic and extrinsic camera parameters are then recovered by finding a parameter set θ (e.g. $\theta = (f_x, f_y, c_x, c_y, ...)$) and $\mathbf{t}$ that satisfies

$$\theta, \mathbf{t} = \underset{\theta,\mathbf{t}}{\operatorname{argmin}} \sum_i (\pi(\theta, \mathbf{X}_i, \mathbf{t}) - \mathbf{x}_i)^2. \tag{2.12}$$

Different calibration procedures exist to find the minimum of this objective. For my work, I mostly used the standard calibration routines in OpenCV [25], an open source computer vision library. In some cases, self-written or other routines were employed (cf. Chapter 4).

2.1.2 Light Transport

The Render Equation

The observed radiance at a given location $\mathbf{x}$ on the image sensor is equivalent to the total radiance that is transported onto this location from the surroundings. Using the pinhole camera model, this is the radiance $L(\mathbf{x}, \mathbf{r})$ received from the pixel ray direction $\mathbf{r} = (p_x, p_y, f)$. If we ignore volumetric effects such as light scattering on smoke or fog, then the radiance received is equivalent to the radiance $L(\mathbf{X}, \omega_o)$ transmitted from the first

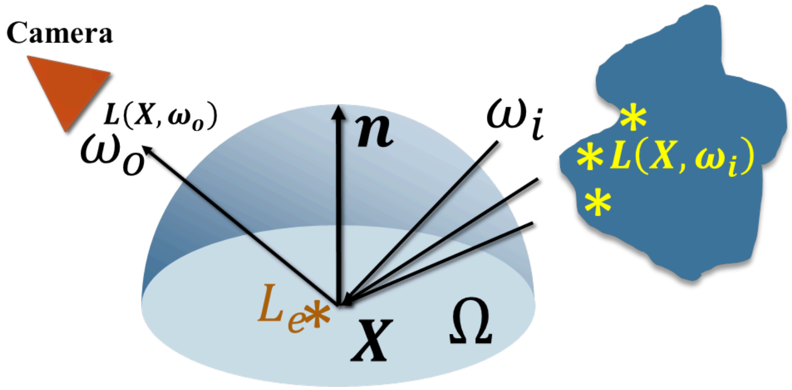


Figure 2.7: Illustration of the Rendering Equation. The amount of light observed in direction ω_i corresponds to the sum of surface emissivity L_e and total amount of incoming light $L(\mathbf{X},\omega_\mathbf{i})$ reflected in direction ω_i.

visible surface point $\mathbf{X}$ in direction $\omega_o = \frac{-\mathbf{r}}{||\mathbf{r}||}$.

This quantity is governed by the render equation [92] (cf. Figure 2.7)

$$L(\mathbf{X},\omega_\mathbf{o}) = L_e(\mathbf{X},\omega_\mathbf{o}) + \int_\Omega f_r(\mathbf{X},\omega_\mathbf{i},\omega_\mathbf{o})L(\mathbf{X},\omega_\mathbf{i})(\omega_\mathbf{i}\cdot\mathbf{n})d\omega_\mathbf{i}, \tag{2.13}$$

where

- $\mathbf{n}$ is the surface normal at $\mathbf{X}$,
- $\omega_\mathbf{i}$, $\omega_\mathbf{o}$ are viewing and outward direction,
- $L_e(\mathbf{X},\omega_\mathbf{o})$ is the radiance emitted by $\mathbf{X}$ in direction $\omega_\mathbf{o}$,
- $L(\mathbf{X},\omega_\mathbf{i})$ is the radiance received at $\mathbf{X}$ from direction $\omega_\mathbf{i}$,
- Ω is the half sphere above the surface,
- $(\omega_\mathbf{i}\cdot\mathbf{n})$ is the geometric attenuation of the incident light due to the surface being at an angle to the incident light and finally,
- $f_r(\mathbf{X},\omega_\mathbf{i},\omega_\mathbf{o})$ is the bidirectional reflectance distribution function (BRDF) that describes the fraction of incident light that is reflected towards $\omega_\mathbf{o}$.

The Radiance L is a power density with the unit $\mathbf{Wm}^{-2}\,\mathrm{sr}^{-1}$. The total amount of light received on a pixel corresponds to this radiance integrated over the pixel surface and the solid angle under which the observed surface appears. Due to the pinhole approximation used here, where pixel surface and subtended angle are infinitesimally small, the observed pixel intensity is proportional to $L(\mathbf{X},\omega_\mathbf{o})$.

The rendering equation expresses the conservation of energy: The radiance observed from under a certain viewing direction from a surface is the sum of the emitted radiance in that direction as well as the total amount of incoming light that is reflected into this direction. It is in essence a geometric optics approximation to Maxwell equations that govern classical electrodynamics [92].

Note that this is the simplest form of the rendering equation. It can be extended to model spectral effects, translucent materials etc. A full treatment of physically based rendering techniques can be found in [145]. As the observed radiance at any one point in space depends on all other surfaces in the scene, the equations can only be solved using Monte-Carlo methods such as path tracing [92] or finite-element methods such as Radiosity [65].

As a final note, it should be mentioned again that the light transport model used here completely ignores the wave nature of light and therefore cannot describe diffraction effects. Methods in computational electromagnetics [43] do exist that directly solve the scalar [91] or vectorial [127] Maxwell equations. Diffraction effects (For visible and near infrared light that are considered) are relevant at spatial scales that are much smaller than the scale of the issues encountered in this work. Therefore they are not considered in the following treatment.

BRDFs

The BRDF is a material specific function that describes the surface reflectance of an object. As it also accounts for light absorption by the material, it is not a distribution in a probabilistic sense: The integral over Ω does not have to yield one. The only requirement made is that no additional light is 'created'.

$$\int_{\Omega} f_r(\mathbf{X}, \omega_{\mathbf{i}}, \omega_{\mathbf{o}})(\omega_{\mathbf{i}} \cdot \mathbf{n}) d\omega_{\mathbf{o}} \leq 1 \, \forall \omega_{\mathbf{i}}. \tag{2.14}$$

Other properties that real world BRDFs have are Helmholtz reciprocity, i.e.

$$L(\mathbf{X}, \omega_{\mathbf{i}}, \omega_{\mathbf{o}}) = L(\mathbf{X}, \omega_{\mathbf{o}}, \omega_{\mathbf{i}}), \tag{2.15}$$

which means that camera and light source can be interchanged without changing the observed intensity. Furthermore, it is required that the BRDF is positive:

$$f_r(\mathbf{X}, \omega_{\mathbf{i}}, \omega_{\mathbf{o}}) \geq 0. \tag{2.16}$$

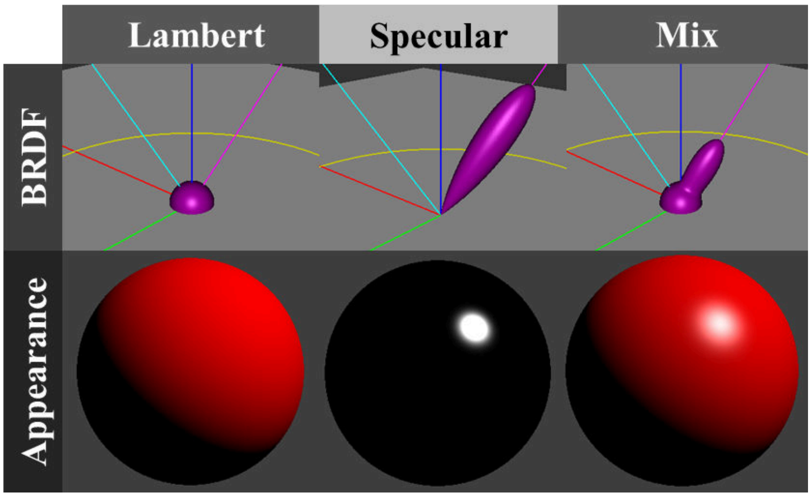


Figure 2.8: Example BRDFs. **Top row**: 3D polar plots of different BRDFs for a fixed light angle (light blue line). Purple line marks the mirror direction. The radial component of the purple surface describes the amount of light reflected into the corresponding direction. **Bottom row**: Appearance of a sphere with corresponding BRDF lit by a point light source. **From left to right**: Diffuse material with isotropic BRDF. Specular material where the light is roughly reflected in mirror direction and finally a material that contains both diffuse and specular components.

BRDFs can either be measured by using a variety of different setups [165, 109] or alternatively, obtained by approximating real BRDFs to varying degree using analytical models. Examples of different analytical BRDFs and the resulting surface appearance are illustrated in Figure 2.8.

Lambertian Materials The majority of vision algorithms assume that the world consists only of Lambertian materials, i.e. all surfaces are perfect diffuse reflectors. For non-emitting materials this can be described as

$$f_r(\mathbf{X}, \omega_\mathbf{i}, \omega_\mathbf{o}) = c, \tag{2.17}$$

for some valid constant c. Equation (2.13) then simplifies to

$$L(\mathbf{X}, \omega_\mathbf{o}) = c \int_\Omega L(\mathbf{X}, \omega_\mathbf{i})(\omega_\mathbf{i} \cdot \mathbf{n}) d\omega_\mathbf{i}. \tag{2.18}$$

Note that the observed intensity of $\mathbf{X}$ no longer depends on the viewing angle ω_o. This property of Lambertian surfaces is the basic assumptions made for most depth imaging techniques. If

$$L(\mathbf{X}, \omega_\mathbf{i}) \propto \delta(\omega_\mathbf{i}, \hat{\omega}), \tag{2.19}$$

is assumed additionally, i.e. parallel light falling from a single direction $\hat{\omega}$, then Equation (2.18) further simplifies to

$$L(\mathbf{X}, \omega_\mathbf{o}) \propto c \cdot (\hat{\omega} \cdot \mathbf{n}). \tag{2.20}$$

This is the main lighting model used in shape from shading [148] techniques. (cf. 2.2). Similarly, for $\hat{\omega} = \omega_\mathbf{o}$ we obtain the lighting model for Time-of-Flight reconstructions.

Perfect Mirrors The other extreme is defined by

$$f_r(\mathbf{X}, \omega_\mathbf{i}, \omega_\mathbf{o}) = \delta(\omega_\mathbf{i} - H(\mathbf{n})\,\omega_\mathbf{o}), \tag{2.21}$$

where $H(v) = I - 2vv^\mathrm{T}$ is the Householder transform that reflects a vector on a plane. A surface described by such a material appears as a perfect mirror. Fortunately, perfect mirrors rarely occur in real life scenarios as in this case. Without additional reasoning it is not possible to extract any information about the surface geometry using vision.

Other Models Most real life materials lie somewhere between these two extremes of perfectly diffuse and perfect mirror materials. While part of the light is reflected diffusely in all directions, another part is reflected along directions close to the mirror direction. Microfacet BRDF models such as Cook-Torrance [38] or Ward [184] can model rough specular reflections by assuming a microscopic distribution of normals at any one point in space. Cook-Torrance additionally accounts for Fresnel reflections, which cause strong specular reflections at grazing angles ($\omega_\mathbf{i} \cdot \mathbf{n} \ll 1$). That is even true for materials like unpolished wood or cardboard, which are traditionally considered to be prototype diffuse materials.

2.1.3 Digitization

Light transport and camera geometry define how the world is projected as an intensity distribution over the whole sensor. The digitization process converts the continuous signal into a digital one that is then stored on the computer. The first step of digitization occurs when the continuous signal is **sampled** by measurements made on the pixel grid. High frequency variations of the intensity above the Nyquist frequency are lost due to the Nyquist-Shannon sampling theorem. If the signal contains such high frequency components, aliasing effects can deteriorate the image if no optical low pass filter is placed in front of the sensor. At the pixel, the collected light is converted into a voltage which is then measured. Due to the particle nature of light, the observed intensity is subject to shot noise, which can be modeled by a Poisson distribution. For large number of photons this can be approximated by a Gaussian distribution with standard deviation $\sqrt{I}$, if I is the signal strength observed. The read-out process and thermal effects in the sensor cause additional errors summarized as a dark signal, which contributes most noticeably at low intensities. The gain and the dark signal can be different for different pixels (effects known as Dark signal non-uniformity and photo-response non-uniformity), leading to a fixed pattern noise in the observed image. This has to be accounted for during calibration.

Finally, the analogue voltage measured in the pixel is then **quantized** at discrete intervals and stored as an integer value which leads to a quantization error. Further information on image digitization and sensor characterization can be found in [85, 48, 49, 84].

2.2 Depth Imaging Techniques

Depth imaging is concerned with recovering geometry from images. Due to the projective nature of imaging systems that leads to a loss of information, depth imaging systems either make use of multiple independent measurements or have to make strong prior assumptions on the type of scene observed. I will first introduce different classes of 3D imaging techniques before giving a detailed description of the two methods under consideration in this work. A full overview of existing methods can be found in [85].

A first distinction has to be made between volume imaging techniques and depth imaging. While volume imaging techniques reconstruct volumetric densities of an absorbing material using tomography or similar approaches, depth imaging is concerned with reconstructing the surfaces present in a scene. This work is concerned with the latter class, which is also often called 2.5D imaging as the depth information gained can be represented in a 2D image grid or a mesh. Depth imaging techniques can be categorized into passive methods that only rely on the natural scene illumination as well as active techniques that illuminate the scene with a coded light source. Another classification that can be made is between triangulation based methods, Time-of-Flight (ToF) imaging and purely radiometric approaches.

Triangulation based approaches include passive techniques such as (multi-view) stereo, light field imaging, structure from motion or depth from focus/defocus[2] as well as active techniques that fall under the term of structured light scanning. These techniques all have in common that they measure depth by directly or indirectly measuring the angle under which a world point appears at two or more different locations. These two angles, together with the spatial relationship between the two points of measurement (called baseline in passive stereo), define a triangle whose third corner corresponds to the desired 3D location. As the baseline is mostly known (save for structure from motion techniques), the main challenge is to find corresponding points in the different measurements that define the two angles.

ToF techniques measure the round trip time of light actively emitted to the scene and back to the sensor. With the speed of light, this round trip time can then be converted into a geometric distance. LIDAR systems mostly use a single sensor that scans the surroundings in order to obtain the geometry of scenes. ToF cameras on the other hand resolve the incoming light spatially by putting it through a camera lens system. If the round-trip time is measured directly, the reconstruction process is fairly straightforward, while modulation based systems commonly applied in ToF cameras require a demodulation step.

The final class of techniques is formed by shape from shading. Here, the light sources are either controlled or an assumption is made on the lighting conditions such as a parallel light with known direction. Additionally, the objects to be imaged are assumed to be composed of Lambertian materials. Under these conditions, the observed intensity (or shading) of the objects is proportional to the angle between surface normal and source of light. The reconstruction process then is concerned with finding a geometry that has normals which are consistent with the measured shading. Newer approaches [12] also try to estimate the ambient lighting while still retaining the Lambertian material assumption.

2.2.1 Passive Stereo

A complete treatment of passive stereo can be found in [75] and [85]. In the following, I will summarize parts relevant to this thesis.

[2] A treatment of depth from focus as a triangulation technique can be found in [161].

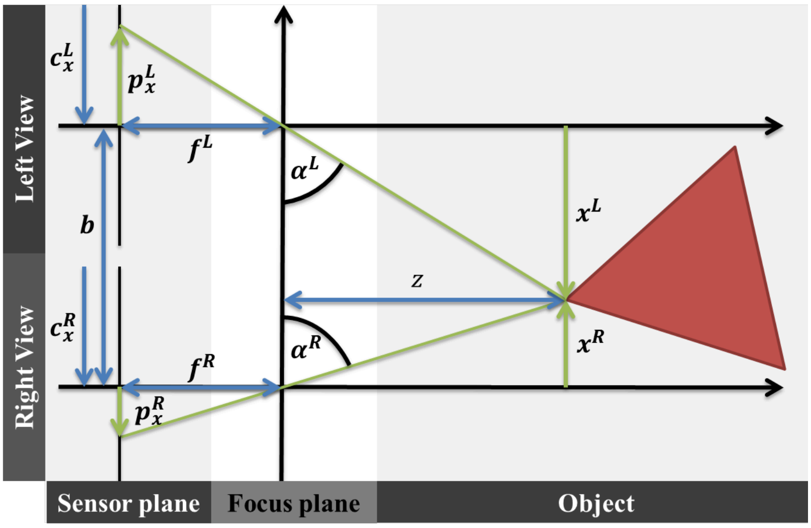


Figure 2.9: Passive Stereo in Two Dimensions. The 3D location X^L in the coordinate frame of the left camera is reconstructed by triangulation: The locations of the projections in the left and right views (p_x^L resp. p_x^R) implicitly define two angles (α^L and α^R) which, together with the known baseline b define a triangle with height z (distance from camera).

Working Principle

The working principle of passive stereo is depicted in Figure 2.9. Let L and R be two pinhole cameras with identical internal parameters ($f^L = f^R = f$, $\mathbf{c}^L = \mathbf{c}^R = \mathbf{c}$) that are placed side by side and are separated by a horizontal baseline b. Let $\mathbf{X^L}$ be some point in the coordinate system of L. The projections of $\mathbf{X^L}$ in L and R are:

$$(p_x^L, p_y^L) = (f\frac{x^L}{z^L} + c_x, f\frac{y^L}{z^L} + c_y), \tag{2.22}$$

and with $\mathbf{X}^R = \mathbf{X}^L - (b, 0, 0)$

$$(p_x^R, p_y^R) = (f\frac{x^L - b}{z^L} + c_x, f\frac{y^L}{z^L} + c_y). \tag{2.23}$$

The displacement between these two points is

$$\mathbf{d} = (p_x^L, p_y^L) - (p_x^R, p_y^R) = (\frac{f \cdot b}{z^L}, 0). \tag{2.24}$$

The horizontal displacement $d_x = f \cdot b/z$ is inversely proportional to the distance of the camera plane from the object while no vertical displacement is observed. This horizontal displacement is called disparity $d = d_x$.

Parameter Estimation

The goal of stereo matching methods is to measure the disparity for each pixel i in L, yielding a disparity map $\mathbf{D} = \{d_i\}$. At the same time, this is the main challenge as we have to infer that two pixels belong to the same location by using only the observed intensities. To make this problem tractable, the majority of techniques in literature make a Lambertian world assumption, which, as discussed above, makes a surface appear the same regardless of the viewing angle. Under this assumption, $I^L(p_x, p_y)$ and $I^L(p_x - d, p_y)$ are the same for the correct disparity d. This is formally expressed in terms of photo consistency between left and right image, where for each pixel (p_x, p_y) a $\hat{d}$ is sought such that

$$\hat{d} = \underset{d}{\operatorname{argmin}} \left(I^L(p_x, p_y) - I^R(p_x - d, p_y) \right)^2 . \tag{2.25}$$

Note that this is just the simplest way to formulate the problem. Refined techniques may penalize the color difference in another manner to be more robust towards effects in real camera systems such as different camera gains.

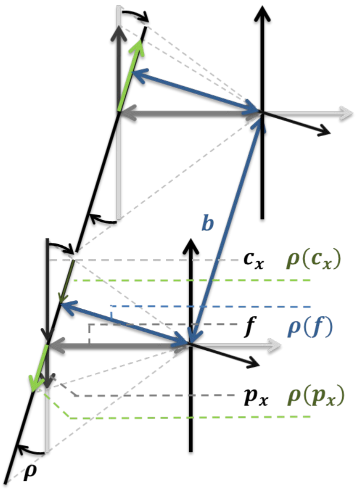


Figure 2.10: Illustration of Stereo Rectification. By virtually rotating the camera sensor using the rectification transform ρ without changing the camera centers, the two sensor planes can be aligned. Standard stereo formulas are applicable using the rectified camera extrinsics $\rho(f)$ and $\rho(c)$. Rectification also transforms the original pixel locations p_x.

Calibration and Rectification

In practice, the two stereo heads can rarely be aligned perfectly for the above formulas to be directly applicable. Also, with radial distortions, the mapping is no longer as straightforward. The basic idea to tackle this is to pre-process the data such that the resulting data behaves like a horizontally aligned pair of pinhole cameras. The first step is to find the actual relative pose between the left and right camera frames via calibration. After warping the images to undistort the images, a so-called rectification transform is applied to each image that virtually rotates the sensor plane, such that the new sensor planes are virtually aligned. An illustration of how this process works is given in Figure 2.10. The two new images resulting from the rectification process behave like a standard horizontal baseline stereo pair with respect to measuring disparity. A detailed treatment of image rectification can be found in [75].

Discussion

Being a modality **closest to human perception**, passive stereo undeniably has exerted a strong fascination on researchers ever since the early days of computational vision [117]. Being a passive sensing technique, it depends on **available scene illumi-**

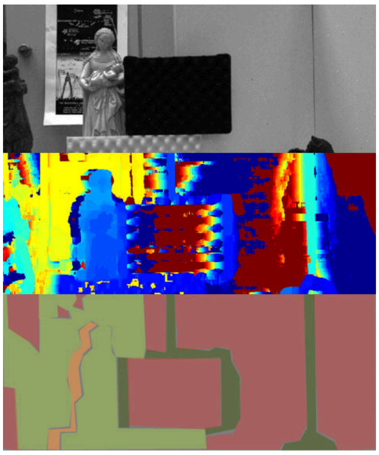


Figure 2.11: Stereo without Regularization. **Top**: Left Image. **Middle:** Disparity Map using 5 × 5 Blockmatching. Colormap ranges from red (far away objects) over yellow (distance to the wall) to blue (close by objects). **Bottom**: Light green: Correct disparity estimation. Dark green: Edge fattening (correct disparity in center). Yellow: Occlusions. Red: Textureless surfaces.

nation. For the same reason it is highly **energy-efficient**[3]. Due to the **widespread availability** of high resolution digital cameras, stereo matching is also arguably the depth imaging technique that can be set up most easily. This also results in depth maps with a **large effective lateral resolution** and the depth resolution can simply be adjusted by changing the baseline[4]. At the same time, the **computational complexity** for recovering a depth map increases with larger images, as for each pixel in the left image all pixels along a line in the right image have to be compared. Furthermore, the stereo matching problem is **severely under-constrained** such that the solution of the problem is highly ambiguous in practice (cf. Figure 2.11). This is especially true for **untextured surfaces** and **highly repetitive textures** (e.g. a chessboard), where many possible solutions minimize Equation (2.25).

To handle the ambiguity, pixel aggregation or another kind of spatial regularization [159] has to be employed, which increases the computational complexity tremendously. These strategies will be discussed in Section 2.3.3. Regularization itself has to be applied carefully, as it can introduce additional model violation errors. Such errors are often observed at **depth edges**, which violate the smooth world assumption frequently encoded in the regularization. While techniques such as data driven weighting [19] or truncated potentials [174] exist, they also introduce additional algorithm parameters which may be challenging to set in such a way that the methods generalize well.

The final class of issues arise due to violations of the photo consistency assumption. One case where this happens is when regions visible in the left image are **occluded** in the right one. A common approach to handling this is by left-right consistency [58] checking or by explicit visibility reasoning [186]. The other case is caused by **specular surfaces**, which can violate the photo consistency assumption between images. This subject will be revisited later in Chapter 6.

2.2.2 Continuous Wave Time of Flight Imaging

This Section is in parts based on the book chapter [106] co-authored by me. While other designs for ToF imagers exist, I

[3] Which also explains why the majority of day-active land animals rather rely on eyes for than on Time-of-Flight measurements (e.g. echo location).

[4] https://xkcd.com/941/

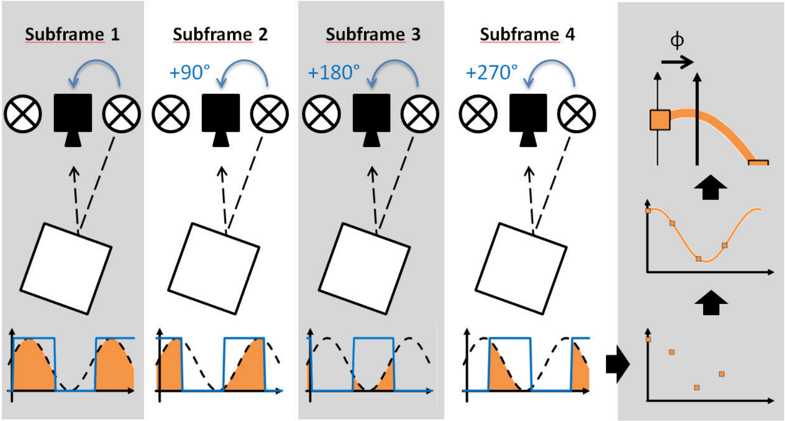


Figure 2.12: Working Principle of a Time of Flight Camera. The light sources are placed symmetrically around the camera to approximate a point light source in the camera center. The modulated light source emits light that is reflected off the scene and arrives at the sensor (black dashed sine function) with a phase shift Φ compared to the input signal. This phase shift corresponds to the distance traveled. The phase shift is obtained by first computing the correlation (orange) between the incident signal with a rectangular signal at different fixed phase shifts (blue) and then estimating the analytical parameters that explain these measurements.

limit myself to Continuous Wave Intensity Modulation (CWIM) based sensors, as this is the working principle of the majority of cameras available today. Besides that, the camera utilized in this thesis also operates on this principle. For a complete presentation of ToF imaging please refer to [67] and [74].

Working Principle

The working principle of Continuous Wave Time of Flight is illustrated in Figure 2.12. The active illumination emits a light with sinusoidally modulated intensity with frequency f_m:

$$e(t) \propto \sin(2\pi f_m t). \tag{2.26}$$

This signal is reflected from the scene back to the camera. The basic imaging properties of a ToF camera are the same as any other camera such that the signal received in each pixel $\mathbf{x}$ corresponds to the reflected light from a certain position $\mathbf{X}$. Assuming the light source to be in the camera center and the distance between camera and focal plane to be negligible, the distance traveled by the reflected light is $2r = 2||x||$ and the round trip time of the signal is $t = 2r/c$ with the speed of light c. The received signal can be modeled as

$$i_r(t) = \frac{g}{\pi} + a\sin(2\pi f_m t + \frac{4\pi f_m}{c} * r), \tag{2.27}$$

where

- g is the reflected ambient light and the DC component of the modulated light,
- a is the reflected amplitude of the modulated light and

- $\frac{4\pi f_m}{c} * r =: \phi$ is the phase shift between the emitted and the reflected signal.

We re-parametrize i_r and e as a function of $\tau = 2\pi f_m t$. Let

$$h(\tau) = \Theta(e(\tau)), \tag{2.28}$$

with Θ being the Heaviside step function, be a rectangular signal with frequency f_m. The correlation function between r and h is

$$\begin{aligned} I^T(\theta) = h(\tau+\theta) \bigotimes i_r(\tau) = \int_0^{2\pi} i_r(\tau) h(\tau+\theta) d\tau \quad &(2.29)\\ = \int_\tau^{\theta+\pi} i_r(\tau) d\tau \quad &(2.30)\\ = \frac{g}{\pi} \cdot \theta - a\cos(\phi+\tau)\Big|_\theta^{\theta+\pi} \quad &(2.31)\\ = g + a\cos(\theta+\phi). \quad &(2.32) \end{aligned}$$

The phase shift ϕ is determined by estimating the parameters of this correlation function by $i = 1, ..., N-1$ samples obtained sequentially in hardware using a Photonic Mixer Device, a special type of pixel sensor [164, 172]. For the camera used in this thesis, the PMD Camcube 3, $N = 4$ and the samples are obtained for $\theta_i = i\frac{\pi}{2}$:

$$I^{T,i} = g + a\cos(i\frac{\pi}{2} + \phi). \tag{2.33}$$

Parameter Estimation

The unknowns that have to be estimated per pixel in Equation (2.33) are g, a and ϕ. Given 4 measured values $I^{T,i}$ the least squares problem can be stated as

$$(\hat{g}, \hat{a}, \hat{\phi}) = \operatorname*{argmin}_{g,a,\phi} \sum_i^4 \left(I^{T,i} - \left(g + a\cos(i\frac{\pi}{2} + \phi) \right) \right)^2. \tag{2.34}$$

The closed form solution to this problem is

$$\begin{aligned} \hat{g} &= \frac{1}{4}\sum_{i=1}^4 I^{T,i}, &(2.35)\\ \hat{a} &= \frac{1}{2}\sqrt{(I^{T,3} - I^{T,1})^2 + (I^{T,0} - I^{T,2})^2}, &(2.36)\\ \hat{\phi} &= \operatorname{atan2}((I^{T,3} - I^{T,1}, I^{T,0} - I^{T,2}). &(2.37) \end{aligned}$$

A derivation can be found in [157]. The distance r can be computed from ϕ as

$$r = \frac{c}{4\pi f_m}\phi. \tag{2.38}$$

Note that the solution $(\hat{g}, \hat{a}, \hat{\phi})$ is determined up to a multiple of 2π as

$$g + a\cos(i\frac{\pi}{2} + \phi) = g + a\cos(i\frac{\pi}{2} + (\phi + 2\pi n))\,\forall n. \qquad (2.39)$$

The range

$$r_{amb} = \frac{c}{4\pi f_m}2\pi = \frac{c}{2f_m} \qquad (2.40)$$

is called the disambiguity range of the camera and defines the maximum range at which the camera can be operated without ambigous depth estimates. Finally, it should be noted that r is a radial depth. The location $\mathbf{X}$ for measurements in pixel (p_x, p_y) can be computed from r as

$$\mathbf{X} = r \cdot \frac{(p_x, p_y, f)}{||(p_x, p_y, f)||}. \qquad (2.41)$$

Calibration

ToF cameras have to be intrinsically calibrated like any other camera, as the intrinsic parameters need to be known to compute world coordinates of the surfaces estimated in each pixel. Additionally, the depth data has to be calibrated to account for different systematic errors ToF systems suffer from. The causes will be discussed in the next section. The general approach towards calibration is to estimate correction functions l_r that are applied on top of the closed form solution given by Equation (2.37) to obtain the corrected solutions:

$$\begin{aligned} G_c &= l_r^G(\hat{G}), && (2.42)\\ A_c &= l_r^A(\hat{A}), && (2.43)\\ \phi_c &= l_r^\phi(\hat{\phi}). && (2.44) \end{aligned}$$

The correction functions range from look up tables [193] to more elaborate model fits [163]. For the experiments in this work, I eitger used the factory calibration provided by the manufacturers or a linear approximation of the factory method.

Discussion

Due to the availability of a closed form solution (Eq. (2.37)) and the computation of $I^{T,i}$ in hardware, depth estimation using CWIM ToF is **fast**[5] and **spares computational resources**.

[5] With the PMD Camcube 3, up to 50 frames per second can be achieved.

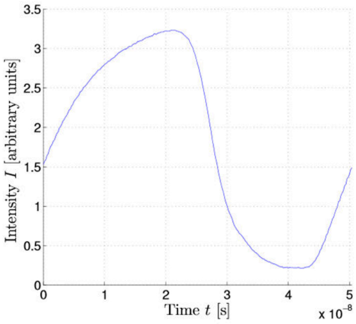


Figure 2.13: Measured non-sinusoidal modulation of the Camcube 3. (Figure courtesy M. Schmidt [163]).

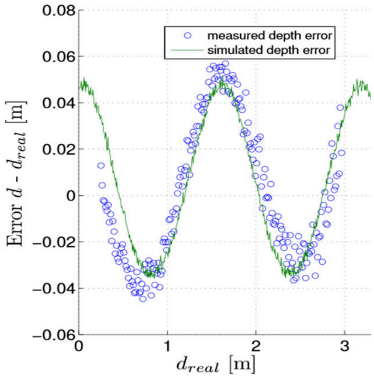


Figure 2.14: Periodic systematic deviation from true depth due to modulation errors. (Figure courtesy M. Schmidt [163])

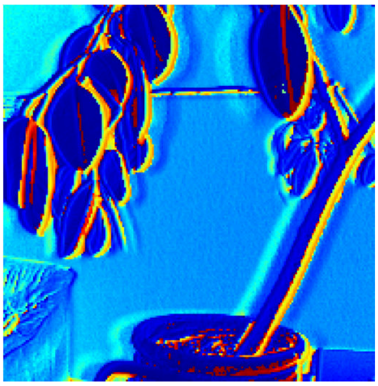


Figure 2.15: Motion Artifacts (yellow and red regions as well as ghost image) visible in ToF depth maps caused by camera movement. (Image Joint work with J.-M. Gottfried[66]))

Unlike passive stereo, ToF imaging also works on **textureless scenes** as well as **in the dark** and unlike structured light systems, ToF imagers can also be utilized **outdoors** to some extent as current cameras contain additional circuitry to suppress ambient light. All these properties make CWIM ToF cameras an attractive alternative for applications such as human computer interfaces, industrial quality control or robot vision.

Yet, ToF imaging does have its own share of issues, which will be discussed in the following. Like passive stereo on textureless surfaces, ToF cameras also have ambiguity issues if the scene extent is larger than the **ambiguity range** (cf. Eq. (2.39) and Chapter 5). Other issues include the **low lateral resolution** of ToF cameras [6] in comparison to standard cameras. The quality of the estimated depth depends on the amount of light reflected as the estimates are more **noisy** in darker areas of the image. Finally, Time-of-Flight cameras are known to suffer from several systematic errors, some of which are scene dependent, while others can be accounted and compensated for by calibration. The main effects are the following:

1. The emitted signal e is not purely sinusoidal in practice (cf. Figure 2.13). Yet, the same closed form solutions are still used. This leads to a characteristic depth dependent oscillation called wiggling error (cf. Figure 2.14) of the estimated depth around the true depth. These deviations can be accounted for to some extent in calibration [163].
2. The four measurements of a pixel are believed to be reflected from the same point in space. As the four measurements are acquired subsequently and not simultaneously, this assumption is violated if the camera is moving or if the scene contains moving objects. These errors are most pronounced depth and intensity edges (cf. Figure 2.15) and can be reduced to some extent using dynamic calibration techniques [163]. Alternatively they can are compensated for by additional modeling of the movement [66].
3. Finally, the signal measured in each pixel is assumed to stem from the single, direct light path between light source over reflected surface to the camera. Due to the finite size of the pixel, the signal is a superposition of all signals received from a certain solid angle of space causing ‘flying pixel errors’ that are most visible at depth discontinuities.

[6] For the camera used in this thesis it is 200×200 px .

Similarly, if global light transport is taken into consideration, the surface not only reflects the light emitted by the camera, but also light reflected from other parts of the scene, which also leads to a superposition. This 'multi-path' error is most pronounced on specular surfaces and on diffuse surfaces at grazing angles. Handling such errors is subject of most recent research on ToF cameras [56, 87].

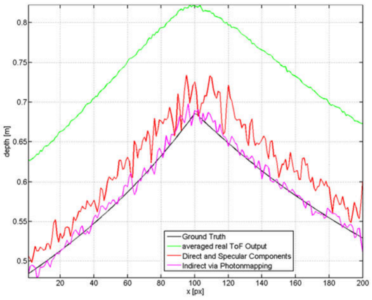


Figure 2.16: Systematic Errors Due to Multi-Path. Black line: Ground Truth profile of a corner. Green line: ToF depth for same profile (not corrected for global offset). Other lines: simulated profiles from [125]. Note the bulge of the measured corner compared to the true corner. (Figure courtesy S. Meister [125])

2.3 Parameter Estimation

Throughout this thesis, I derive models aimed at explaining effects observed in depth imaging systems by means of scene and camera parameters. To obtain these parameters from observations, I make use of existing parameter estimation techniques or present novel strategies based on combinations thereof. Below, I will briefly present the types of optimization[7] problems encountered as well as the methods commonly used to solve them.

2.3.1 Structure of Optimization Problems

In my work, two basic kinds of parameter estimation problems are encountered. **Dense problems** [159, 10, 97, 29] are dealt with in Chapters 3, 5 and 6 and arise when depth or other parameter maps are to be estimated for each pixel of an observed image. Examples already encountered are the disparity map that is estimated by stereo matching, or the radial depth, amplitude and intensity estimation problem of ToF cameras (for which a closed form exists). The other kind of problem is encountered during pose estimation (Chapter 4) and camera calibration where a sparse set of observations is considered. It is known as the **bundle adjustment** problem [179, 191].

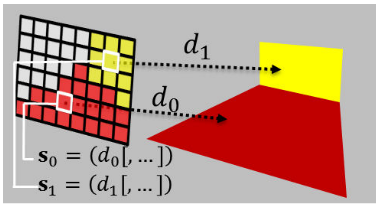

d_1
d_0
$\mathbf{s}_0 = (d_0[, \dots])$
$\mathbf{s}_1 = (d_1[, \dots])$

Figure 2.17: Illustration of Dense Problems. For standard stereo reconstruction, $\mathbf{s_i}$ corresponds to the disparity d_i that is to be estimated in each pixel of the left camera. Similarly, for ToF reconstructions, $\mathbf{s_i} = (g_i, a_i, \phi_i)$ is the vector containing offset, amplitude and phase.

Dense Problems Let $\Omega^D \subset \Omega$ denote the set containing the pixel locations on the image plane of the primary camera (e.g. the left stereo camera or the ToF camera). The basic assumption is that the scene can be described by a parameter vector $\mathbf{s}_i$ defined for each pixel $i \in \Omega^D$ (cf. Figure 2.17). With bold letters, e.g. $\mathbf{S}$, $\mathbf{G}$, etc., I denote the set of all parameters or the set of all parameter components, i.e. $\mathbf{S} = \{\mathbf{s}_i\}$, $\mathbf{G} = \{g_i\}, ..., i \in \Omega^D$.

[7] In the following, 'parameter estimation' is used synonymously with the other terms 'optimization' and 'inference'.

The true parameters $\hat{\mathbf{S}}$ best describe the N observed images $I = \{I^j\}$, $j = 0, ..., N-1$, given some additional constants θ (camera intrinsics, baseline, speed of light etc.) and the measurement model. This is formulated in terms of a function E that has a minimum for $\hat{\mathbf{S}}$, i.e.

$$\hat{\mathbf{S}} = \underset{S}{\operatorname{argmin}}\, E_d(\theta, \mathbf{S}, I). \tag{2.45}$$

In literature, E_d is called the objective function, cost function or energy function[8] of the problem. In many cases Equation (2.45) decomposes into a sum

$$E_d(\theta, \mathbf{S}, I) = \sum_{i \in \Omega_D} E(\theta_i, \mathbf{s}_i, I \setminus I^0), \tag{2.46}$$

where $\theta_i = \theta \cup \{I^0_i, i\}$. I^0 is the image as viewed from the primary camera (over which Ω^D is defined) and I^0_i is a short hand for the intensity at pixel location i, $I^0(i)$. It will become clear why the distinction between I^0 and the other images is made further below. As each term in the sum only depends on one of the unknown $\mathbf{s}_i$, the argmin operation can be estimated independently for each $\mathbf{s}_i$. In such cases, E is frequently also called a cost function or a (generalized) cost volume [152]. Stereo matching (Eq. (2.25)) and ToF reconstruction (Eq. (2.34)) can both be expressed as a dense reconstruction problem (with $I = \{I^L, I^R\}, I^0 = I^L$ for stereo and $I = \{I^{T,j}\}, j = 1, ..., 3, I^0 = I^{T,0}$ for ToF imaging):

$$E_{stereo}(\{i, I^L_i\}, d_i, I^R) = \left(I^L_i - I^R\left(i - (d_i, 0)\right)\right)^2. \tag{2.47}$$

$$E_{ToF}\begin{pmatrix} \{i, I^{T,0}_i\}, \\ \{g_i, a_i, \phi_i\}, \\ \{I^{T,1}, I^{T,2}, I^{T,3}\} \end{pmatrix} = \sum_{j=0}^{3}\left(I^{T,j}_i - \left(g_i + a_i \cos(j\tfrac{\pi}{2} + \phi_i)\right)\right)^2. \tag{2.48}$$

Note that in Equation (2.46) the cost is summed over each pixel in the primary view only. For the stereo cost, it is thus only guaranteed for the primary image $I^0 = I^L$ that every pixel is explained by the model. The other images are mainly utilized to ensure model consistency. Also, if continuous disparities d_i are considered, then some form of interpolation has to be applied to evaluate the cost for non-integer d_i. This kind of "model-centric" approach to parameter estimation is very common in vision and also used in this thesis as it is required to make problems com-

[8] This is due to analogies to energy minimization problems in physics [170].

putationally tractable.

Bundle Adjustment (BA) Let V be a set of different camera views indexed by v. Each view is associated with extrinsic and intrinsic camera parameters $\mathbf{t}_v$ and θ_v as well as the image domain Ω^v. Furthermore, let $\{\mathbf{X}_j\}, j = 0, ..., M-1$ denote a set of M 3D points and $V(j) \subset V$ denote the views in which $\mathbf{X}_j$ is visible.

For BA problems, the measurements are not the intensities observed in each pixel, but rather a sparse set of locations $\mathbf{x}_{vj} \in \Omega^v$ in each image plane on which $\mathbf{X}_j$ projects onto. Note that the $\mathbf{x}_{vj}$ in general are estimated to subpixel accuracy, i.e. the locations are not discrete. The set of locations $\{\mathbf{x}_{vj} | v \in V(j)\}$ belonging to a single $\mathbf{X}_j$ is called a keypoint track or a set of correspondences. Given the $\mathbf{x}_{vj}$, the goal of BA is to jointly estimate 3D locations $\{\hat{\mathbf{X}}_j\}$ as well as intrinsics and camera pose $\{\hat{\theta}_v, \hat{\mathbf{t}}_v\}$ of each view. The minimization problem is

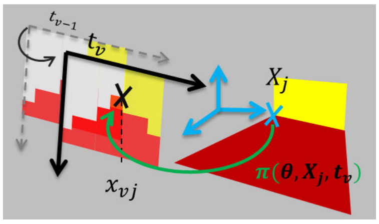


Figure 2.18: Illustration of the Bundle Adjustment(BA) Problem. The goal is to find intrinsic and extrinsic parameters θ and $\mathbf{t}_v$ and in some cases also 3D positions $\mathbf{X}_j$ such that the deviation between

$$\{\hat{\mathbf{X}}_j\}, \{\hat{\theta}_v, \hat{\mathbf{t}}_v\} = \underset{\{\mathbf{X}_j\}, \{\theta_v, \mathbf{t}_v\}}{\operatorname{argmin}} \; E^{BA}(\{\mathbf{X}_j\}, \{\theta_v, \mathbf{t}_v\}), \tag{2.49}$$

with

$$E^{BA}(\{X_j\}, \{\theta_v, \mathbf{t}_v\}) = \sum_{j=1}^{N} \sum_{v \in V(j)} ||\mathbf{x}_{vj} - \pi(\theta_v, \mathbf{X}_j, \mathbf{t}_v)||_2^2, \tag{2.50}$$

where $\pi(\theta_v, \mathbf{X}_j, \mathbf{t}_v)$ is the projection operation defined in Equation (2.11). This energy simply states that the true parameters are the ones that minimize the reprojection error between observed correspondences $\mathbf{x}_{vj}$ and the projections of their corresponding 3D point $\mathbf{X}_j$. This is the full BA problem. If some of the other parameters can be measured in advance, more simplified calibration problems arise. If $\{\mathbf{X}_j\}$ is known, for example when the points belong to a calibration target, we obtain the calibration problem

$$\{\hat{\theta}_v, \hat{\mathbf{t}}_v\} = \underset{\{\theta_v, \mathbf{t}_v\}}{\operatorname{argmin}} E^{BA}(\{\mathbf{X}_j\}, \{\theta_v, \mathbf{t}_v\}). \tag{2.51}$$

If, i addition, all θ_v are known this further simplifies to the extrinsic calibration problem (used for stereo calibration)

$$\{\mathbf{hatt}_v\} = \underset{\{\mathbf{t}_v\}}{\operatorname{argmin}} E^{BA}(\{\mathbf{X}_j\}, \{\theta_v, \mathbf{t}_v\}). \tag{2.52}$$

2.3.2 Optimization Strategies

For both types of problems encountered, model parameters **S** have to be found, which minimize an energy function which in turn penalizes (=has large values for) incorrect parameters. If no closed form solution to the optimum can be obtained, depending on the domain over which **S** is defined, either *discrete* or *continuous* optimization techniques can be considered for parameter estimation. For BA problems, continuous techniques are used whereas both techniques are prevalent for dense problems.

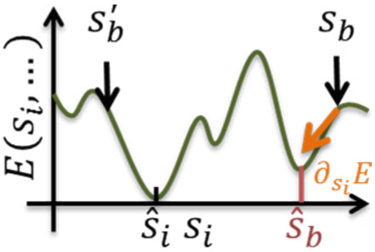


Figure 2.19: One Dimensional Energy Surface. Using local features such as the gradient, continuous optimization finds the closest local minimum. The right initialization (s_b' vs s_b) is required to find the global minimum.

Continuous techniques assume $E(\mathbf{S})$ to be real valued, continuous and differentiable to a certain degree in **S**. Depending on the degree of differentiability, different local features of $E(\mathbf{S})$ such as the Jacobian or Hessian can be used to obtain a direction in which **S** must be changed to further decrease the value of the objective function. While continuous methods yield more accurate results than their discrete counterparts, the outcome strongly depends on the initialization of the methods, since, as a rule of thumb, continuous techniques will converge to the 'nearest' local minimum of the energy function(cf. $\hat{s}_b$ in Figure 2.19). The global minimum can therefore only be found if the objective is convex , i.e has a single local minimum which simultaneously is the global minimum. Unfortunately, only few problems in vision (and none of the ones presented in this thesis) can actually be formulated as a convex problem. One approach to handling the non-convexity is by convexifying the cost (e.g by using scale space approaches [5]). Another approach, as employed in my work, is to choose the initial value sufficiently close to the true minimum, by sampling or some other kind of initialization. Further information on continuous techniques can be found in [136] and [23].

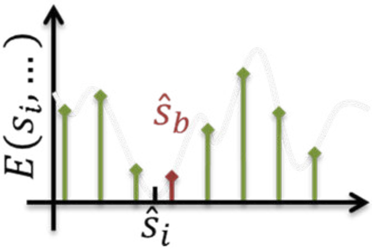


Figure 2.20: Continuous Energy Surface Discretized. By testing all possible values a solution close to the global minimum can be found if the discretization step is fine enough.

Discrete methods, on the other hand, operate on a so-called 'label space', where the continuous domain of **S** is quantized to discrete values, for example when only integer valued disparities are required in stereo matching. While the resulting optimization problems are often NP hard, such methods have the advantage that they can overcome local minima. Integer and combinatorial optimization techniques are applied to solve such problems [134]. The simplest of such techniques is an exhaustive search over all combinations of parameters (Figure 2.20). For dense problems which decompose (cf. Eq. (2.46)), the computational burden can be reduced by doing the grid search for each pixel independently. Line search based methods in stereo [94, 159] are an example for such a strategy.

2.3.3 Regularization

As vision problems are frequently under-constrained, there are often multiple solutions with the same argmin. In these cases, prior knowledge can be added in form of a regularizer R, which is included in the overall objective function. This leads to a regularized total energy function of the form

$$E_d^r(\theta, \mathbf{S}, I) = E_d(\theta, \mathbf{S}, I) + R(\mathbf{S}). \tag{2.53}$$

In many cases, this regularizer only depends on the model parameters $\mathbf{S}$ and some additional constants. More recent research also presents data dependent regularization [86, 110], where the regularization strength is computed from image features. These methods also build on the classic techniques which will be reviewed here. R frequently has the form

$$R(\mathbf{S}) = ||\Gamma(\mathbf{S} - \mathbf{S_0})||_p^p, \tag{2.54}$$

where $\mathbf{S_0}$ contains prior information on the desired location of $\mathbf{S}$, $||.||_p$ is the p-norm and Γ is an (often linear) operator encoding the desired structure of the solution (e.g. a discrete differential operator encoding smoothness). To give an intuition, three special cases will be discussed here. For $\Gamma = \mathbf{1}$ we obtain

$$R(\mathbf{S}) = ||\mathbf{S} - \mathbf{S_0}||_p^p = \sum_{i \in \Omega_D} |s_i - s0_i|^p, \tag{2.55}$$

we demand a solution that is close to a **prior solution $\mathbf{S_0}$** (cf. Figure 2.21). If E_d decomposes into individual terms (as in the case for many dense problems), the regularized objective does so as well.

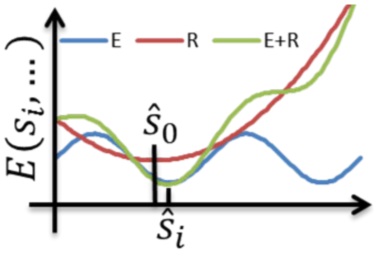


Figure 2.21: Regularization: A regularization term (red) is often required to make the otherwise ambiguous minimum of the cost function (blue) unambiguous (green). In this case there is prior information that favors a solution near $\hat{s_0}$.

For dense problems with $\mathbf{S_0} = \mathbf{0}$, Γ often encodes the **spatial smoothness** of the desired solution, e.g. that the differences between neighboring pixel parameters shall be small. A simple example for such a spatial regularization is the objective

$$E_d^r(\theta, \mathbf{S}, I) = \sum_{i \in \Omega_D} E(\theta_i, s_i, I \setminus I^0) + \sum_{j \in N(i)} ||s_i - s_j||_p^p, \tag{2.56}$$

where $N(i)$ is the set containing the locations of neighboring pixels (cf. Fig 2.22). For $p = 2$, small large changes in parameters incur huge additional costs, while small changes are hardly penalized. Therefore small smoothly varying parameter maps without discontinuities are favored. L1 regularization (p=1) does not penalize large changes as heavily, thus allowing for discontinuities in the parameter map. Another kind of implicit spatial regularization is **cost aggregation**, which is employed in patch based methods such as block matching or patch match [11]. Here, the individual terms in Eq. (2.46) are aggregated over a pixel neighborhood

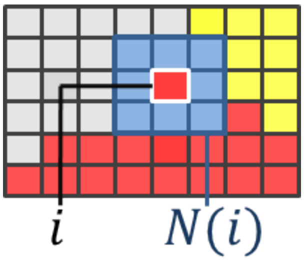


Figure 2.22: Illustration of the neighborhood $N(i)$ of pixel i. In some cases larger neighborhoods may also be considered.

$$E^{agg}(\theta_i, s_i, I \setminus I^0) = \sum_{j \in N(i)} E(\theta_j, s_i, I \setminus I^0), i \in \Omega_D. \tag{2.57}$$

Such an aggregation can - to some extent - be interpreted as a strong local regularization of parameters over the whole patch. Yet, unlike methods to solve Eq. (2.56), patch based approaches can still be evaluated independently. This also means that the extent of the regularization depends on the neighborhood size.

In Eq. (2.56), the total energy of the objective depends on the values of all parameters, yielding a high dimensional optimization problem. Optimization methods again depend on the domain of the solution: Variational [154, 33, 190] or diffusion based [160, 144] methods are used to solve continuous problems, while techniques from graphical model inference [97, 24, 130, 108] are used for discrete problems.

2.4 Summary

This chapter presented an overview of theory, notation and background information required for the following parts of this thesis. I commenced by reviewing the basic aspects governing image formation. I put an emphasis on the pinhole camera model, as it is used in all chapters, and on light transport since the derivation in Chapter 6 is based on this. Deviations of real camera images from this model are either corrected for (e.g radial distortions, non-parallel stereo) or are considered to be negligible. The subsequent section gave an introduction to the two depth-imaging techniques considered: ToF cameras and passive stereo. Here, the focus was on the basic working principle and derivation of the measurement model before I give an overview of the properties of both systems. The least square formulation for parameter estimation will be revisited throughout the next chapters. The final section of this chapter was committed towards the parameter estimation process itself. As this thesis makes use of existing or combinations of existing methods, the goal of the last section was to give an intuition on the types of problems encountered as well as to impart an overview of the resulting optimization techniques. In the following chapters, it will be shown how specific models naturally lead to certain optimization strategies.

3

ToF - Stereo Fusion

This following chapter is based on my work previously presented in [131] and [133].

3.1 Motivation

WILL THERE EVER BE one depth sensor to rule them all? While this will hopefully be true one day, all current depth sensing modalities fall short of obtaining this title. In the previous chapter (cf. Section 2.2), I presented the working principle of passive stereo and ToF imaging and discussed the strengths and weaknesses thereof. Summarizing it, passive stereo works well in presence of scene texture and, due to the availability of mega-pixel cameras, has the potential to produce high resolution depth imagery. Conversely, there are issues a) at occlusion boundaries, b) when the textures are ambiguous or c) when no texture is present at all. Additionally, the parameter estimation process is computationally demanding due to the large solution space, moreso if global optimization techniques are considered. Time-of-Flight (ToF) imaging, on the other hand, delivers depth images at high frame rates independent of surface texture, but at the cost of a lower resolution, sensor noise and systematic errors.

The two techniques considered differ considerably in the areas where they excel or fail. Therefore, it appears natural to combine them to create a more reliable system.

3.2 Contributions and Outline

In the following chapter, I present a system that produces high resolution depth reconstructions by combining ToF and stereo data. The data-fusion is implemented on the GPU enabling fast parameter estimation at interactive rates. It differs from existing work by the usage of dense per pixel confidence measures to guide the reconstruction. Results validating the method are presented on scenes with and without reference data for quantitative evaluation. To this end, I present one of the first publicly available reference datasets for purposes of benchmarking ToF-stereo fusion methods. Finally, from a theoretical perspective, I investigate how the methods presented here as well as in literature relate to the model that suggests itself by combining the raw measurement models of the individual modalities (cf. Eqs. (2.47) and (2.48)). It turns out that all existing techniques can be derived from this “full-model” by a series of approximations and modifications motivated by different assumptions on the measurement errors. To the best of my knowledge, this is the first time that the ToF-stereo fusion problem has been formulated in such a way.

The remainder of this chapter is organized as follows: After discussing the related work in Section 3.3, I continue with a discussion of the camera system I set up as well as the design considerations leading to this setup (Section 3.4). In Section 3.5, I then establish the full measurement model for ToF-stereo fusion and derive simplifications which are used in literature and in the subsequent section (Section 3.6). Here, the confidence based fusion approach as well as the utilized optimization strategy are presented. After discussing experiments and results in Section 3.7, I finally conclude the chapter with a summary and outlook on open questions and future work in Section 3.8.

3.3 Related Work

The full related work on data fusion for 3D reconstruction covers a wide range of topics including combinations of multiple color views (multi-view) [166], fusion of stereo and depth from defocus [149, 175], ToF and a single camera [141, 83], multiple ToF cameras [31], sonar and stereo [120] or structure from motion and depth imaging [135]. A presentation of the related work in this widest sense is out of scope of this chapter. Therefore, I will limit myself to ToF-stereo fusion techniques with a focus on a high level classification of methods. To achieve this, I first present the general pipeline employed by the majority of methods before giving an overview of how the methods differ. An in-depth treatment of these methods can be found in [133].

Pipeline Most fusion systems differ mainly in how the data is merged once it has been brought into the same reference frame. Figure 3.1 illustrates the basic pipeline employed by the majority of methods. After choosing a specific camera setup, the intrinsic parameters for the stereo and ToF cameras have to be estimated, i.e. focal length, principal point and distortion coefficients. Next, the spatial relationship (rotation and translation) between the three cameras has to be found by means of a pairwise stereo calibration or alternatively by joint calibration together with the depth calibration of the ToF camera.

For the ToF camera, additionally a depth calibration has to be undertaken to account for the systematic errors described in Section 2.2.2. This is either done using standard ToF calibration techniques [106], or done jointly with the extrinsic calibration of the stereo system [193, 41, 162]. After applying preprocessing steps to clean up the ToF data (i.e. to reduce effects by noise pixels), the images must be brought into the same coordinate frame by means of rectification and reprojection. Finally, data fusion involves one or more of the following steps:

- The ToF depth and the output of a stereo algorithm are computed individually and then fused.
- The ToF data is used as an initial guess and to reduce the search space for subsequent stereo refinements.

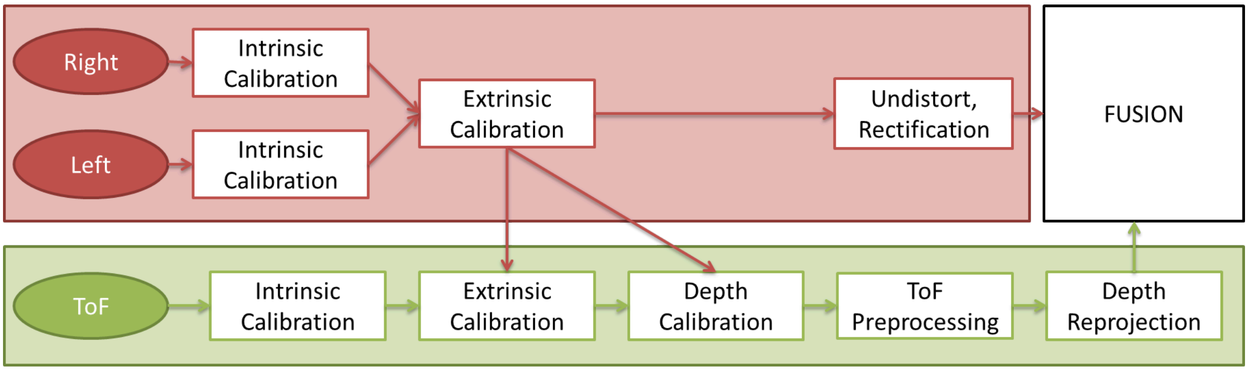


Figure 3.1: Fusion Pipeline. The majority of related ToF-stereo methods differ from each other in the way the data is merged after bringing ToF and Stereo data into the same reference frame.

- The depth reconstruction algorithm uses both stereo and ToF costs as data terms.

The techniques additionally differ from each other in the kind of regularization techniques that have been applied.

ToF-Stereo Fusion Methods Fusion techniques can either be categorized in terms of the time of fusion or in terms of the optimization strategy that is employed. **Time of fusion** refers to the point in the fusion pipeline, where the individual modalities are fused together. Late fusion techniques [42, 104] first compute depth maps from ToF and stereo independently before they combine these two sources. On the other hand, early fusion methods on the other hand (encompassing all other methods mentioned here) use the ToF depth estimates to initialize and regularize the stereo matching procedure. In the work presented, I additionally introduce the 'symmetric early fusion' problem, where parameters have to be found that simultaneously explain ToF and stereo raw measurements. The derivation of an inference strategy for this problem is subject to future work. As the majority of techniques belong to the early-fusion class, grouping the methods in terms of the **optimization strategy** is more practical. Following [159], which makes a similar taxonomy for stereo algorithms, the methods can be grouped in local and global methods.

Local methods [104, 68, 13, 72, 41, 188, 131, 14] tend to be faster and parallelizable but cannot cope with locally erroneous or ambiguous data. They are often based on a line search that is guided by the ToF data. [68] applies a hierarchical stereo matching algorithm directly on the remapped TOF depth data without considering uncertainties. [104, 72] compute confidences in the ToF image and let stereo refine the result in regions with low confidence. The latter are similar to the method presented here, but only use binary confidence maps based on the ToF amplitude image and therefore only sparsely use stereo information. Instead, the local method proposed here uses the information of both data sources in the form of data fidelity measures to guide the fusion process. As such it is most similar to [41]. The main difference to [41] is the choice of data term that allows reconstruction without having to visit the full cost volume. *Global methods* [54, 71, 193, 192, 194, 98, 131, 155, 59, 42, 169] add the ToF information as an additional data term in a global energy functional, which is then jointly optimized. While the depth maps obtained are smoother due to the usage of prior informa-

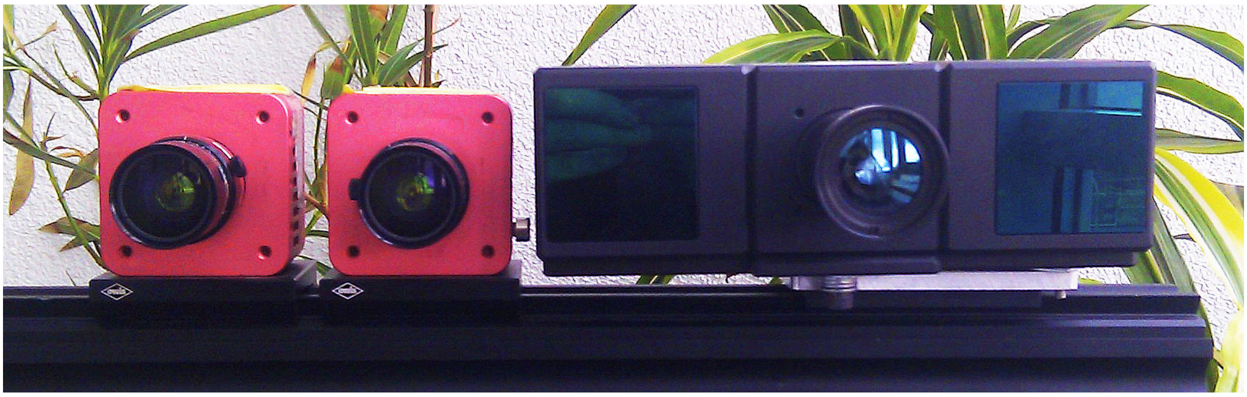


Figure 3.2: Camera setup. The stereo subsystem (Red cameras) consists of two Photon Focus MV1-D1312-160-CL-12 with Linos Mevis-C lenses at 25mm/1.6, 1312x1082px. The ToF camera (Black camera)is a PMDTech Camcube 3, 200x200px .

tion/regularizers, this is at the cost of additional computational complexity. These global techniques can be further grouped depending on the framework that was chosen for optimization. [71, 193, 192, 194, 169] employ different probabilistic inference techniques on (discrete) *graphical models*. Being discrete, the accuracy is limited to pixel level. Moreover, such methods do not scale well; the stereo images are of lower resolution then the ones considered in this work.

[155, 131] formulate the problem in a *variational* framework. Since it is a continuous technique, the problem of initialization arises. [155] relies on a scale space approach, while the work we presented in [131] depends on the close initialization using the local technique presented here. The last sub-group of the global methods [54, 98, 59, 42] contains those which use *other non-local* optimization strategies such as semi-global matching [77] or seed growing [32]. The work presented here is not primarily concerned with regularization but with the data term used for matching and as such can be utilized in combination with any regularization frame work. As evidence of algorithm performance using global methods, I present results from [131] that are based on joint work on combining the data terms presented here and variational regularization.

3.4 Camera Setup

3.4.1 Acquisition Setup

The camera setup is depicted in Figure 3.2. It consists of two high-resolution cameras[1] (L, R) and a low resolution ToF cam-

[1] Photon Focus MV1-D1312-160-CL-12 with Linos Mevis-C lenses at 25mm/1.6 (≈ 35° FOV), 1312x1082px.

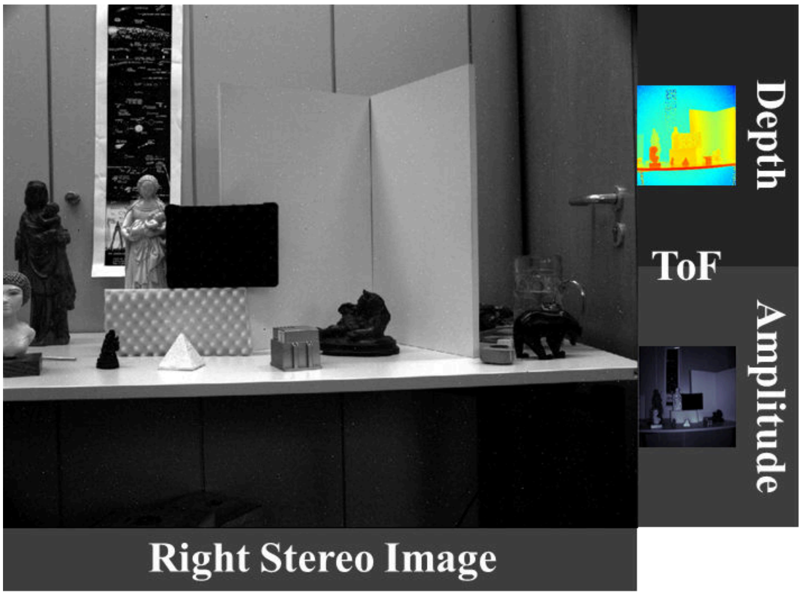


Figure 3.4: Input: Time of Flight data (200x200), One of the stereo images (1312x1082). Images are depicted to scale.

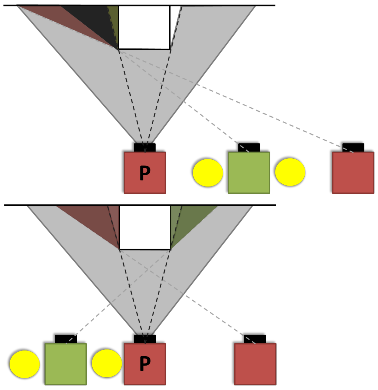


Figure 3.3: Camera Placement for Stereo-Centric Fusion. **P** marks the primary (left stereo) camera. **Top:** A symmetric camera placement causes the region of occlusion between ToF and Stereo to overlap, thus creating areas that remain occluded in the left view (dark gray area). **Bottom** Assymetric camera placement allows the ToF subsystem to account for areas not visible by the right stereo camera such that the complete field of view of the left camera can be reconstructed.

era[2] (T). The stereo camera was connected to the acquisition PC equipped with a frame grabber card[3] via Camera Link; the ToF camera via USB. The stereo subsystem was synchronized and triggered by the frame-grabber in hardware. Synchronization of the ToF camera was achieved by operating the camera in software trigger mode and implementing a call-back triggered by the frame-grabber API. Example frames acquired from this system are displayed in Figure 3.4.

Camera placement is a simple, yet important aspect of the camera setup often overseen in literature. The majority of methods presented in Section 3.3 place the ToF camera in between the two stereo heads while still using the left stereo camera as the reference frame. As illustrated in Figure 3.3, this creates some areas in the left image that cannot be reconstructed as they are occluded both in the right and in the ToF image. In the setup presented, the ToF camera is positioned such that the regions of occlusion do not overlap. This enables depth estimation in all areas of the primary camera.

3.4.2 Calibration

Intrinsic and extrinsic calibration of camera parameters was done using the OpenCV calibration modules [25] and a checker board target. For the ToF camera, the intensity images of the target

[2] PMDTech Camcube 3 with standard 12.8 mm lenses (40° FOV), 200x200px.

[3] SiliconSoftware microEnable IV.

were bicubicly upsampled by a factor of 5 and then used as input to the calibration methods as this yielded the best results in terms of reprojection error.

A total of 50 target images were acquired for each camera individually and for each pair of cameras for pairwise stereo calibration. The reprojection error was around a tenth of a pixel for intrinsic calibration of the stereo and the ToF camera[4] and a fifth of a pixel for the pairwise stereo calibration. The left and right images were compensated for radial distortion and additionally rectified, while ToF image was only compensated for radial distortion. For depth calibration, the inbuilt correction of the ToF camera was applied. It should be noted that by using the stereo calibration routine, the extrinsic calibration between the ToF and stereo subsystem is less accurate than the extrinsic calibration between the individual stereo heads. This is in general an open problem as it leads to alignment errors during reprojection of the ToF depth on the stereo head. It will be discussed further in Section 3.8 and Chapter 4.

3.5 Modelling ToF Stereo Fusion

3.5.1 Least Squares Formulation of ToF-Stereo Fusion

Let $V = \{L, R, T\}$ denote local 3D coordinate systems of the left, right and ToF cameras and let W denote the world coordinate frame. With $\mathbf{X}^c, c \in V \cup \{W\}$, I refer to a 3D position in terms of one of these frames. The camera frames are connected to the world frame via transformations $(R^v, s^v) = \mathbf{t}^v, v \in V$. These are obtained via calibration[5]. When used as a mapping, $\mathbf{t}^v$ converts a 3D point from one representation into another (cf. Eq. (2.10)), e.g.

$$
\begin{aligned}
\mathbf{X}^L &= \mathbf{t}^L(\mathbf{X}^W), && (3.1)\\
\mathbf{X}^W &= \mathbf{t}^{L^{-1}}(\mathbf{X}^L), && (3.2)\\
\mathbf{X}^L &= \mathbf{t}^L(\mathbf{t}^{R^{-1}}(\mathbf{X}^R)) = \mathbf{t}^{RL}(\mathbf{X}^R). && (3.3)\\
& && (3.4)
\end{aligned}
$$

Each of the views $v \in V$ also defines a projective mapping

$$
\pi^V : \mathbb{R}^3 \rightarrow \Omega^V, \pi^V(\mathbf{X^W}) = \pi(X^W, \theta^v, \mathbf{t}^v), \qquad (3.5)
$$

[4] 1/10 of the ToF image size prior to upsampling

[5] for the stereo centric fusion model that is presented in the next section

with θ^v denoting the intrinsics of camera V obtained via calibration and Ω^v describing the image planes of the respective cameras (cf. Eq. (2.11)). Similarly, I define

$$\begin{aligned} r^v(\mathbf{X^W}) &= ||t^v(\mathbf{X^W})||, v \in V, & (3.6)\\ z^v(\mathbf{X^W}) &= (0,0,1) \cdot t^v(\mathbf{X^W}), v \in V, & (3.7)\\ d^L(\mathbf{X^W}) &= \frac{bf}{z^L(\mathbf{X^W})}. & (3.8) \end{aligned}$$

The first two denote the radial and z distances of a point $\mathbf{X^W}$ from the camera centers, while the last mapping is the z depth converted into a disparity, given a baseline b and a focal length f of the stereo system given by the left and right camera. The locations on the image plane on which the pixels sample the intensity are defined by the discrete set $\Omega^v_d \subset \Omega^v$. Finally, by

$$I^L(\mathbf{x}),\ I^R(\mathbf{x}) \text{ and } I^{T,i}(\mathbf{x}),\ i = 0, ..., 3, \tag{3.9}$$

I define the color or intensity at location $\mathbf{x} \in \Omega^v$ on the image plane with bilinear interpolation if $(x, y) =: \mathbf{x} \notin \Omega^v_d$:

$$I(x,y) = \begin{bmatrix} \lceil x \rceil - x \\ x - \lfloor x \rfloor \end{bmatrix}^T \begin{bmatrix} I(\lfloor x \rfloor, \lfloor y \rfloor) & I(\lfloor x \rfloor, \lceil y \rceil) \\ I(\lceil x \rceil, \lfloor y \rfloor) & I(\lceil x \rceil, \lceil y \rceil) \end{bmatrix} \begin{bmatrix} \lceil y \rceil - y \\ y - \lfloor y \rfloor \end{bmatrix}. \tag{3.10}$$

$\lfloor\rfloor$ and $\lceil\rceil$ denote floor and ceiling operations that return the closest points on the pixel grid. For the ToF image, the additional superscript i indexes the individual sub-frames of the ToF image. As a short hand, I also use $I^v, v \in V$ to refer to the whole image.

To derive the least squares formulation of the ToF-stereo reconstruction problem, let us consider a point $\mathbf{X}^W$ on a surface that is visible in all three cameras. Assuming a purely Lambertian world, this point is a solution for the stereo least squares problem (cf. Eq.(2.25))

$$\begin{aligned} \mathbf{X}^W &\overset{!}{=} \underset{\mathbf{X}}{\operatorname{argmin}}\, E_{stereo}(\mathbf{X}) & (3.11)\\ &= \underset{\mathbf{X}}{\operatorname{argmin}} \left(I^L\left(\pi^L(\mathbf{X})\right) - I^R\left(\pi^R(\mathbf{X})\right)\right)^2. & (3.12) \end{aligned}$$

This, as a reminder, encodes the photo consistency constraint that is applicable to Lambertian surfaces.

Simultaneously, together with the correct ToF amplitude a and intensity g, this same point also is a solution to the ToF least squares problem (cf. Eq. (2.34)):

$$(g, \mathbf{X}^W, a) \overset{!}{=} \underset{(g,\mathbf{X}^W,a)}{\operatorname{argmin}} E_{ToF}(g, \mathbf{X}, a), \tag{3.13}$$

with $E_{\text{ToF}}(g, \mathbf{X}, a) =$

$$\sum_{i}^{4} \left(I^{T,i}(\pi^T(\mathbf{X})) - \left(g + a\cos(i\frac{\pi}{2} + r^T(\mathbf{X})\frac{4\pi f_m}{c})) \right) \right)^2 . \tag{3.14}$$

These considerations lead to the *full symmetric model* for ToF stereo fusion for a single point:

$$(g, \mathbf{X}, a) = \underset{(g,\mathbf{X},a)}{\operatorname{argmin}} E_{stereo}(\mathbf{X}) + \lambda\, E_{ToF}(g, \mathbf{X}, a). \tag{3.15}$$

Here, λ is a coupling factor that accounts for radiometric differences between the ToF and stereo sensors.

3.5.2 Camera-Centric Fusion

In camera-centric reconstructions, the world coordinate system is set to coincide with one of the camera systems, also called the primary camera frame:

$$(\exists! P \in V) P \overset{!}{=} W. \tag{3.16}$$

Additionally, the assumption is made, that the scene geometry can be described adequately by one of the following scalar fields that is sampled discretely on the primary image plane:

$$\begin{aligned} \mathbf{Z} &= \{z_i\}, i \in \Omega_d^P, & (3.17)\\ \mathbf{R} &= \{r_i\}, i \in \Omega_d^P, & (3.18)\\ \mathbf{D} &= \{d_i\}, i \in \Omega_d^P. & (3.19) \end{aligned}$$

At each pixel location i, these maps describe the geometry in terms of a z-depth map, a radial depth map (as delivered natively by the ToF camera) or a disparity map d_i (as estimated natively by a stereo system). Together with the camera intrinsics, P defines a mapping from d_i to the corresponding 3D location $\mathbf{X}^P(d_i)$ (same hold true for z_i and r_i). The superscript P is omitted in the following if the primary reference frame is evident. Depending on the choice of P, different variants of Eq. (3.15) can be derived.

Stereo-centric fusion For $P = L$ we obtain the stereo centric fusion model that is most commonly utilized in literature. Here, the geometry is described by a disparity map $\mathbf{D}$ such that the following simplifications can be made:

$$\begin{aligned} \pi^L(\mathbf{X}(d_i)) &= i, & (3.20) \\ \pi^R(\mathbf{X}(d_i)) &= i - (d_i, 0) \overset{!}{=} i - d_i. & (3.21) \end{aligned}$$

The last term is just for notational purposes. This simplifies

$$\begin{aligned} E(d_j, g_j, a_j) = &\Big(I^L(j) - I^R(j - d_j)\Big)^2 \\ &+ \lambda \sum_{i=0}^{3} \Big(I^{T,i}(\pi^T(\mathbf{X}(d_j))) \\ &- \Big(g_j + a_j \cos\Big(i\frac{\pi}{2} + r^T(\mathbf{X}(d_j))\frac{4\pi f_m}{c}\Big)\Big)\Big)^2 \\ &= E_{stereo}(d_j) + \lambda\, E_{ToF}(d_j, g_j, a_j). \qquad (3.22) \end{aligned}$$

Note that d, g and a are all defined in terms of the left camera frame, i.e. $j \in \Omega_d^L$.

ToF-centric fusion Similarly, for $W = P = T$ we obtain the ToF centric fusion model which is formulated in terms of the radial distance $\mathbf{R}$ from the ToF camera center ($j \in \Omega_d^T$).

$$\begin{aligned} E(r_j, g_j, a_j) = &\Big(I^L\Big(\pi^L(\mathbf{X}(r_j))\Big) - I^R\Big(\pi^R(\mathbf{X}(r_j))\Big)\Big)^2 \\ &+ \lambda \sum_{i}^{4} \Big(I^{T,i}(j) - \Big(g_j + a_j \cos(i\frac{\pi}{2} + r_j\frac{4\pi f_m}{c}))\Big)\Big)^2. \end{aligned} \qquad (3.23)$$

3.5.3 Approximations and Modifications

In practice, stereo centric fusion is much more common than ToF-centric, which is why the following approximations will be made for this case. Figure 3.5 displays these approximations:

Quadratic Approximation of $\mathbf{E_{ToF}}$ Often the ToF camera does not deliver the raw channels $I^{T,i}$, but directly outputs closed form depth amplitude and intensity maps. To obtain a least squares problem in terms of these, a Taylor approximation of the problem around the closed form solution of E_{ToF} in Eq. (3.22) in terms of the closed form solution $\mathbf{s_i}^{ToF} = (g_i^{ToF}, d_i^{ToF}, a_i^{ToF})'$

can be made.

$$E = E_{stereo}(d_i) + \lambda\, E_{ToF}(\mathbf{s}_i) \tag{3.24}$$

$$\approx E_{stereo}(d_i) + \lambda\, (\mathbf{s}_i - \mathbf{s}_i^{ToF})' H_{E_{ToF}(\mathbf{s}_i^{ToF})} (\mathbf{s}_i - \mathbf{s}_i^{ToF}), \tag{3.25}$$

with $\mathbf{s_i} = (g_i, d_i, a_i)^T$. The constant term does not contribute to the computation of the argmin of the objective and is therefore omitted. The linear term vanishes as $\mathbf{s}_i^{ToF}$ is the global minimum of E_{ToF}. Note thatλ still only encodes radiometric differences of the sensor. Furthermore, observe that $\mathbf{s_i}^{ToF}$ is defined in the left stereo camera frame. Therefore, to obtain this representation from the ToF camera output, the radial depth, intensity and amplitude images obtained by the ToF camera have to be reprojected into the left camera frame. Details of how this is done for the proposed methods can be found in Section 3.6.1. This quadratic approximation still maintains the correlation of parameters close to the ToF minimum. Yet, the periodicity of the ToF solution is lost.

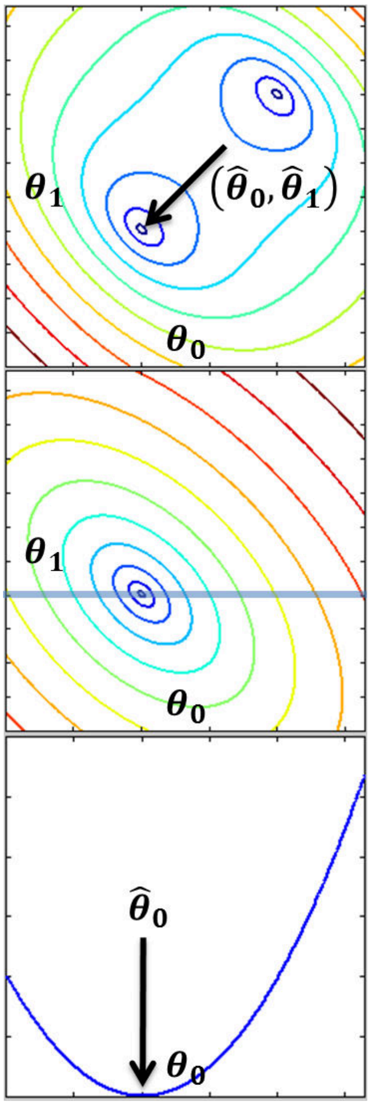


Figure 3.5: Illustrations of the approximations made on a 2D toy example. **Top** Contour plot of a complicated two dimensional objective function **(cf.** E_{ToF} **in Eq.** (3.22)**)**. Red resp. blue lines indicate iso-levels of large resp. small values of the objective function. The global mimimum is indicated with $(\hat{\theta}_0, \hat{\theta}_1)$. **Middle** Quadratic approximation of the objective replacing the energy surface with a quadratic function **(cf.** E_{ToF} **in Eq.** (3.25)**)**. The approximation only holds in areas close to $(\hat{\theta}_0, \hat{\theta}_1)$. **Bottom** Further simplification by only considering the objective along a single dimension (blue line in middle image). The correlation structure (tilt of the quadratic function) is lost **(cf.** E_{ToF} **in Eq.** (3.29)**)**.

"Early Fusion Model" In general, the off diagonal elements $H_{E_{ToF}(\mathbf{s}_i^{ToF})}$ are non-zero in Eq. (3.25), leading to mixed terms:

$$\begin{aligned} E(g_i, d_i, a_i) = & E_{stereo}(d_i) + \lambda \cdot \\ & \Big(c_0 (g_i - g_i^{ToF})^2 \\ & + c_1 (d_i - d_i^{ToF})^2 \\ & + c_2 (a_i - a_i^{ToF})^2 \\ & + c_4 (a_i - a_i^{ToF})(d_i - d_i^{ToF}) \\ & + c_6 (g_i - g_i^{ToF})(d_i - d_i^{ToF}) \\ & + c_7 (a_i - a_i^{ToF})(g_i - g_i^{ToF}) \Big), \end{aligned} \tag{3.26}$$

with constants $c_0, ..., c_6$. The next simplification is made by running the minimization only over the d_i and setting

$$g_i = g_i^{ToF}. \tag{3.27}$$

$$a_i = a_i^{ToF}. \tag{3.28}$$

This further simplifies Eq. (3.26) and leads to

$$E(d_i) = E_{stereo}(d_i) + \mu\, (d_i - d_i^{ToF})^2, \tag{3.29}$$

with a constant $\mu = c_1\lambda$. Note that this constant now also accounts for the different units of the two terms. Eq. (3.29) is commonly called the *early fusion* model in literature[6]. The ToF depth estimate (reprojected and converted into a disparity) is essentially used as per pixel prior to the stereo matching algorithm employed (cf. Eq. (2.55)). Compared to the full quadratic approximation, the correlation structure between amplitude, intensity and disparity is ignored.

"Late Fusion Model" The *late fusion model* is obtained in a similar manner by replacing $E_{stereo}(d_i)$ in Equation (3.29) with a second order Taylor approximation. Due to a similar (and simpler) derivation as above, the objective can be written as

$$E(d_i) = \nu\,(d_i - d_i^{stereo})^2 + \mu\,(d_i - d_i^{ToF})^2, \tag{3.30}$$

where d_i^{stereo} is the solution to the stereo matching problem alone and ν is a constant comparable to μ. This amounts to a weighted mean of ToF and stereo measurements.

Beyond Least Squares The following again deals with the early fusion model. However, the treatment does extend naturally to the other models presented.

An assumption made during the derivation above is that all measurement errors are unbiased and normally distributed. In reality, this does not have to hold - especially for ToF estimates, which are subject to several systematic errors such as wiggling or multi path. Enforcing a quadratic penalty term for the ToF depth would therefore also bias the final reconstruction. A common approach to handling non-Gaussian noise and systematic errors is to replace the quadratic ToF term in Eq. (3.29) with a general loss function Φ that may depend on additional parameters β:

$$E(\mathbf{D}) = \sum_{i\in\Omega_d^L} E_{stereo}(d_i) + \mu\,\Phi_\beta(d_i, d_i^{ToF}). \tag{3.31}$$

For

$$\Phi^{LSQ}(d_i, d_i^{ToF}) = \beta(d_i - d_i^{ToF})^2, \tag{3.32}$$

we obtain the same least squares data term as in Eq. (3.29). Other common loss functions used in literature are the truncated

[6] Due to the fact that the ToF camera is used as a black box range camera in most papers presented.

costs of the form

$$\Phi^{LSQ}(d_i, d_i^{ToF}) = \begin{cases} \beta_0 \left| d_i - d_i^{ToF} \right| & \text{if } \beta_0 \left| d_i - d_i^{ToF} \right| < \beta_1 \\ \beta_1 & \text{else,} \end{cases} \tag{3.33}$$

for parameters $\beta = (\beta_0, \beta_1, p)$. This kind of a loss allows for heavy tailed measurement errors or outliers in the measurement. As a note, different loss functions can be applied to the stereo term as well as we will see in the presentation of the global method in Section 3.6.3. A treatment of the loss functions used in ToF-stereo fusion literature can be found in [133].

3.6 Parameter Estimation

So far, I have presented a general derivation of the basic models used in the majority of ToF-stereo fusion literature. As I mentioned, the difference between the techniques then is in the choice of loss function, level of approximation, how the constants (μ,(ν,) and β) appearing in the models are obtained and in the actual strategy employed to recover the depth maps.

In the following, I present techniques to solve the (robust) stereo centric early fusion model (cf. Eq. (3.31)). Note that I have described these methods previously in [131]. I commence in Section 3.6.1 by describing the reprojection step required to obtain d_i^{ToF}. Subsequently, I discuss image based measures that are used in the following section to approximate the model constants. Finally, in Section 3.6.3, I describe inference strategies employed to recover high resolution depth maps using these confidence measures.

3.6.1 Depth Reprojection

Stereo centric early fusion techniques require the ToF parameter maps to be reprojected onto the left image plane. The location $\mathbf{x}$ on the (continuous) left image plane Ω^L, where the parameters (r_i, a_i, g_i) at pixel $i \in \Omega_d^T$ (ToF frame) reproject onto, is given by

$$\mathbf{x}(i) = \pi^L \left(\mathbf{t}^{TL} \left(\mathbf{X}^T(r_i) \right) \right). \tag{3.34}$$

The amplitude and intensity at this location can simply be copied, while the depth measurement has to be transformed into the disparity space of the left image and thus leading to

$$g_{\mathbf{x}(i)} = g_i. \tag{3.35}$$

$$a_{\mathbf{x}(i)} = a_i. \tag{3.36}$$

$$d_{\mathbf{x}(i)} = d^L \left(\mathbf{t}^{TL} \left(\mathbf{X}^T(r_i) \right) \right). \tag{3.37}$$

Note that$g_{\mathbf{x}(i)}$, $a_{\mathbf{x}(i)}$ and $d_{\mathbf{x}(i)}$ are not defined on the image grid of the left image and sparsely spread over the left image plane. From this sparse reprojection, I obtain a dense sampling on the image grid by means of (linear) interpolation. For $j \in \Omega_d^L$ and $s \in \{g, a, d\}$ this is given by

$$s_j^{ToF} = \sum_{i \in \Omega_d^T} \alpha_{ij} s_{\mathbf{x}(i)}, \tag{3.38}$$

with some interpolating factors α_{ij}. These parameter maps are used as the priors for the early fusion model (cf. Eq. (3.31)).

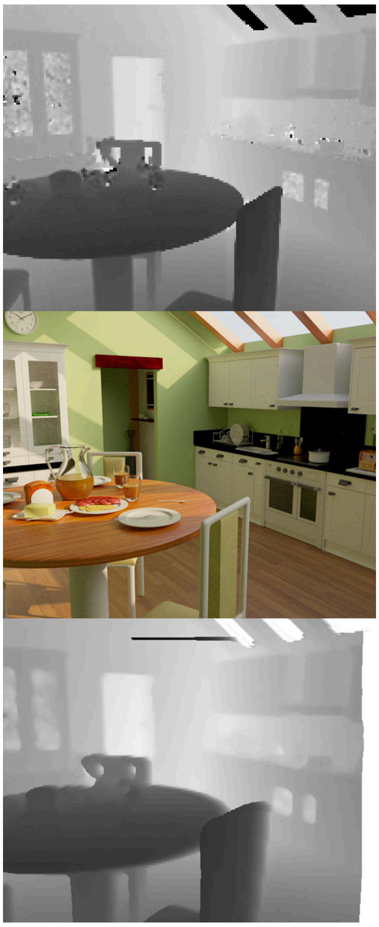


Figure 3.6: Illustration of reprojection and interpolation step on synthetic data. **Top**: ToF range image (dark is near, bright is far). **Middle:** Left color image. **Bottom:** Reprojected depth image. Note the fattening occuring on the right-hand side of object silhouettes. The confidence measures in the next section account for this. N.B. This depth image has to be converted into a disparity map in a subsequent step.

In practice, I implemented the whole reprojection and interpolation process as an OpenGL shader program (cf. Figure 3.6) The ToF depth image is triangulated to create a surface mesh with the vertices corresponding to the $\mathbf{X}^T(r_i)$. The transformed range maps $g_{\mathbf{x}(i)}$, $a_{\mathbf{x}(i)}$ and $d_{\mathbf{x}(i)}$ are stored at each vertex in form of a texture map. The surface is then rendered in the left view with z-buffering yielding the desired parameter maps with interpolation. Other than operating in real-time ($>$ 30 frames per second) on commodity graphic cards[7], using the OpenGL rendering pipeline also has the advantage that, due to the z-buffering, regions in the ToF camera occluded in the left view are automatically removed from the reprojected parameter maps.

3.6.2 Confidence/Uncertainty Measures

With the d_i^{ToF} recovered, the only thing remaining before describing the optimization strategy is the question how the model parameters (μ, and β) are derived.

In principle, they can be computed by measuring the radiometric properties of the camera systems and then by subsequently evaluating the Hessian and other terms that arise in the simplification process. In practice, however, the raw data required for the computations is not always available and the involved calibration and other computations may become rather complex. The methods I present employ a heuristic approach

[7] e.g. GeForce GT 540M

towards obtaining these parameters by means of ad-hoc confidence measures derived from the input images. The idea is based on the observation that in some cases simple image features can be frequently used to predict such system parameters [151]. In the following I present the four confidence/uncertainty measures that are subsequently used. The measures are based on the left stereo image I^L as well as the reprojected ToF parameter maps $A^{ToF} = \{a_i^{ToF}\}$ and $R^{ToF} = \{r_i^{ToF}\}$, $i \in \Omega_d^L$.

ToF Amplitude The quality of ToF depth estimates depends on the amount of (modulated) light reflected from the scene into the camera. Pixels with smaller amplitude will contain more noise than pixels with a large amplitude. I therefore define

$$C^a(i) = \frac{1}{a_i^{ToF}}, \tag{3.39}$$

as the confidence measure that encodes this inverse relationship. Note that this measure ignores the contribution of the intensity image g^{ToF} to the ToF depth confidence [163] and as such must be interpreted as a heuristic.

ToF Range gradient Due to reprojection and upsampling, edges in the reprojected ToF depth image have a high uncertainty. To account for this, I define the measure

$$C^r(i) = ||\nabla_i R^{ToF}(i)||, \tag{3.40}$$

where ∇_i is the discrete (image) gradient operator. Like the measure above, this one has a large value for high uncertainty in the reprojected ToF data.

Left Horizontal Gradient Stereo matching only works in presence of vertical image texture. A large horizontal image gradient (∂_{i_x}) in the left stereo image hence accounts for the situation when stereo matching is likely to work:

$$C^I(i) = ||\partial_{i_x} I^L(i)||. \tag{3.41}$$

Occlusion Map Finally, given the initial disparity map computed by reprojection, it is possible to precompute areas, in which no stereo reconstruction is possible due to the areas being occluded in the right stereo image. This occlusion map is

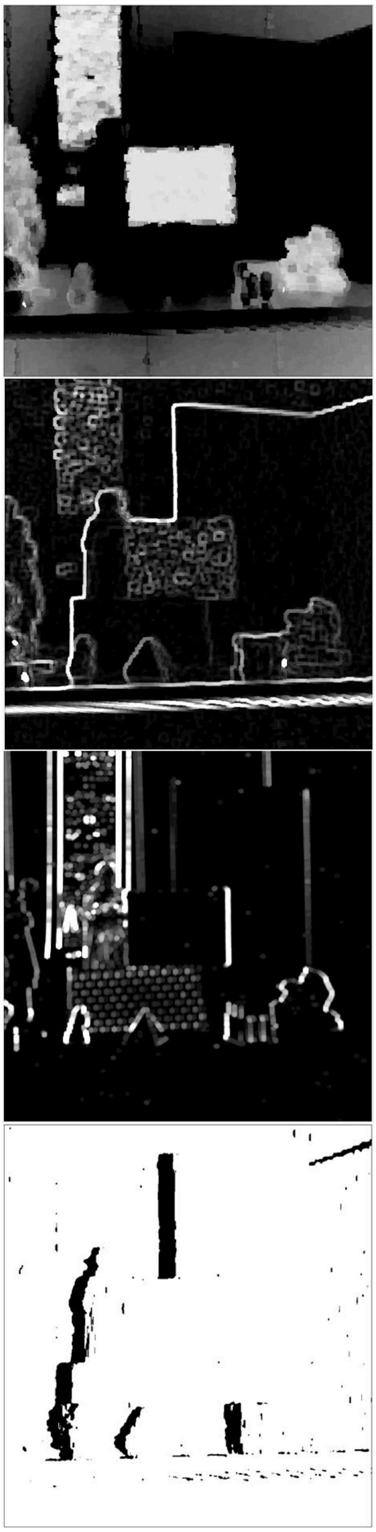


Figure 3.7: Confidence/ uncertainty maps of the scene displayed in Figure 3.4. All values are considered to be normalized to the range $[0, ..., 1]$. **From top to bottom**: C^a, C^r,C^I and C^{Occ}.

defined as

$$C^{Occ}(i) = \begin{cases} 1 & \text{if } \Big(\exists j \in \Omega^L(i)\Big)\; d_j + ||i-j|| > d_i \\ 0 & \text{else} \end{cases}. \qquad (3.42)$$

$\Omega^L(i) \subset \Omega^L_d$ denotes all pixels right of i, and $||i-j||$ denotes the horizontal distance between i and j. The measures are illustrated in Figure 3.7. How these measures can be used in ad-hoc strategies for model inference is presented in the next section.

3.6.3 Optimization Strategies

The following section presents two strategies for solving the ToF-stereo fusion problem. The first strategy is based on cost aggregation and a local grid search (blockmatching) as presented in Section 2.3.3. The second one incorporates an adaptive global regularization term. Both techniques make use of the data terms presented in Section 3.5 and rely on the confidence measures presented in the previous section.

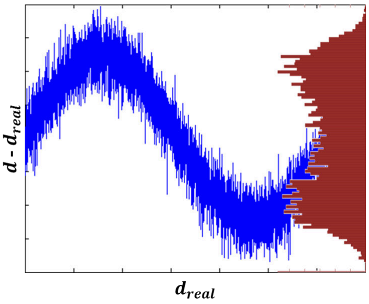


Figure 3.8: Motivation for Loss Used in Local Fusion. **Blue**: Toy example resembling the wiggling error in Figure 2.14. **Red**: Histogram of depth deviations. Due to systematic errors, the ToF depth error is no longer Gaussian. However, it is still bounded.

Local Fusion

The loss functions discussed in Section 3.5.3 still assume no systematic deviation in the data. For ToF cameras, the systematic deviation is bounded in most cases (cf. Figure 2.14), leading to an error distribution that to some extent resembles a uniform distribution (cf. Figure 3.8).

This motivates the usage of a well-loss of the form

$$\Phi^{Well}_{\beta}(d_i, d_i^{ToF}) = \begin{cases} 0 & \text{if } \left|d_i - d_i^{ToF}\right| < \beta \\ \infty & \text{else,} \end{cases}. \qquad (3.43)$$

Using this loss for the ToF term, Eq. (3.31) turns into

$$E(\mathbf{D}) = \sum_{i \in \Omega^L_d} E_{stereo}(d_i) + \mu\, \Phi^{Well}_{\beta_{0,i}}(d_i, d_i^{ToF}) \qquad (3.44)$$

$$= \sum_{i \in \Omega^L_d} E_{stereo}(d_i) + \Phi^{Well}_{\beta_{0,i}}(d_i, d_i^{ToF}). \qquad (3.45)$$

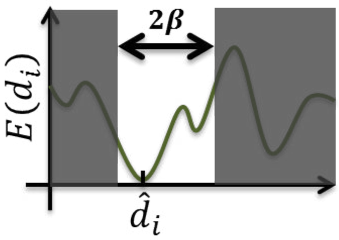


Figure 3.9: Illustration of Eq. (3.45). The stereo solution is in effect constrained to a certain feasable region.

The effect of this objective is illustrated in Figure 3.9. In essence, the (unbiased) stereo matching cost is constrained to lie in proximity of the (biased) ToF solution.

For local fusion, E_{stereo} is replaced with a aggregated version of the stereo cost (cf. Eq. (2.57)):

$$E(\mathbf{D}) = \sum_{i \in \Omega_d^L} E_{stereo}^{agg}(d_i) + \Phi_{\beta_{0,i}}^{Well}(d_i, d_i^{ToF}) \tag{3.46}$$

with

$$E_{stereo}^{agg}(d_i) = \sum_{j \in N(\beta_{1,i}, i)} \left(I^L(j) - I^R(j - d_j) \right)^2 . \tag{3.47}$$

$N(\beta_{1,i}, i)$ denotes a $\beta_{1,i} \times \beta_{1,i}$ patch neighborhood around pixel i with additional parameter $\beta_{1,i}$. Note that$\beta_{0,i}$ and $\beta_{1,i}$ are defined per pixel. They are obtained from the per pixel confidence/uncertainty measures. For $\beta_{0,i}$, we note that it corresponds to the region around the ToF result, where the stereo matching cost is evaluated. This region should be larger in areas we distrust the ToF data and where we are confident in the stereo data. This can be formulated as

$$\beta_{0,i} = C^{Occ}(i) \cdot (\alpha_0, \alpha_1, \alpha_2) \begin{pmatrix} C^a(i) \\ C^r(i) \\ C^I(i) \end{pmatrix} . \tag{3.48}$$

Cost aggregation causes well-known edge fattening effects at object boundaries. To avoid such effects in the proposed system, $\beta_{1,i}$ is also chosen adaptively depending on the depth gradient encoded with $C^r(i)$:

$$\beta_{1,i} = \begin{cases} \gamma_0 & \text{if } C^r(i) > \gamma_1 \\ \gamma_2 & \text{else} \end{cases} . \tag{3.49}$$

This binary selection between window sizes γ_0 and γ_2 is mostly motivated by the fact that it was easier to implement in the existing framework and also gave satisfactory results.

The global parameters $\alpha = \{\alpha_0, \alpha_1, \alpha_2\}$ and $\gamma = \{\gamma_0, \gamma_1, \gamma_2\}$ have to be obtained empirically for a given setup.

With α and γ defined, the only remaining unknown is the disparity map we seek. These are then found using block matching, i.e. exhaustive grid search of Eq. (3.46) for integer valued disparities.

Global Fusion

The global technique was jointly developed with Frank Lenzen, my focus lying on the data term. I will briefly present the relevant aspects of this fusion technique and refer to [131] for further details on implementation.

Here, we considered an objective function of the form

$$E^{TV}(\mathbf{D}) = R(\mathbf{D}) + \sum_{i \in \Omega_d^L} E^1_{stereo}(d_i) + \Phi^1_{\beta_{0,i}}(d_i, d_i^{ToF}), \quad (3.50)$$

with regularizer R and weighted L1 loss for ToF *and* stereo.

$$E^{\beta_{1,i}}_{stereo}(d_i) = \beta_{1,i} |I^L(j) - I^R(j - d_j)|, \quad (3.51)$$

$$\Phi^1_{\beta_{0,i}}(d_i, d_i^{ToF}) = \beta_{0,i} |d_i - d_i^{ToF}|. \quad (3.52)$$

This has the effect that the solution is more robust with respect to outliers in the data. As in the case of local fusion, The parameters $\beta_{1,i}$ and $\beta_{0,i}$ are obtained from the confidence measures above:

$$\beta_{0,i} = C^{Occ}(i) \left((C^a(i)^{-1} - 1)(1 - \beta_{1,i}) \right), \quad (3.53)$$

$$\beta_{1,i} = C^I(i) \cdot C^{Occ}(i). \quad (3.54)$$

Essentially, $\beta_{1,i}$ and $\beta_{0,i}$ are switching variables that continuously choose between the stereo and ToF data term depending on the presence of local texture. If no texture is present *and* the amplitude is low (dark, untextured areas), a case where both ToF and stereo are uncertain, the method down-weights both the data terms and instead relies solely on regularization.

The regularizer R encodes the spatial smoothness of the desired result in terms of first and second order total variation. A somewhat compact way of writing this is

$$R(\mathbf{D}) = \sum_{i \in \Omega_d^L} \mu_0 ||\Psi(\nabla \mathbf{D}(i), \beta_{2,i}, \beta_{3,i})|| + \mu_1 ||\Psi(\text{trace}(H_{\mathbf{D}}(i)), \beta_{2,i}, \beta_{3,i})||, \quad (3.55)$$

with global parameters μ_0 and μ_1 and binary pixel parameters $\beta_{2,i}$ and $\beta_{3,i}$. $\nabla \mathbf{D}(i)$ and $H_{\mathbf{D}}(i)$ are the gradient and Hessian of the parameter map $\mathbf{D}$ computed from finite differences (e.g. by convolution of $\mathbf{D}$). The function Ψ, together with $\beta_{2,i}$ and $\beta_{3,i}$, encodes the adaptive data driven regularization employed here:

$$\Psi(\mathbf{v}, \beta_{2,i}, \beta_{3,i}) = \Psi((v_x, v_y), \beta_{2,i}, \beta_{3,i}) = (v_x \beta_{2,i}, v_y \beta_{2,i}). \quad (3.56)$$

That is, if $\beta_{2,i}$ or $\beta_{3,i}$ is zero, then no regularization (smoothing) is applied at this point. Here, $\beta_{2,i}$ and $\beta_{3,i}$ encode the locations of horizontal and vertical edges in the left image which were obtained using a Canny-like approach to binarize the edge locations [156].

Once all parameters and constants have been obtained, Eq. (3.50) is then solved by means of variable splitting [37, 185] and application of primal dual optimization [33].

3.7 Experiments and Results

3.7.1 Qualitative comparison of ToF, Stereo and the proposed methods

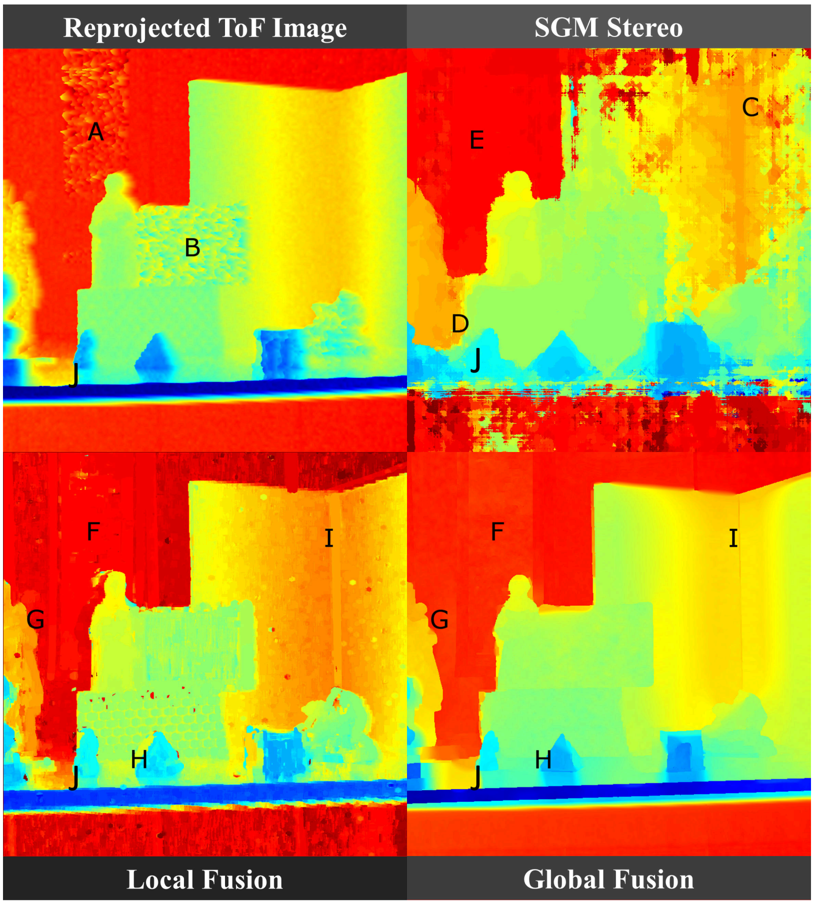


Figure 3.10: Qualitative comparison of the proposed local and global fusion approaches with ToF reprojection and SGM stereo. The color map ranges from red for far away objects over yellow to blue for objects nearby. The letters in the images mark regions that are discussed in Section 3.7.1.

Both proposed methods make use of the CUDA framework an were implemented in C++. Figure 3.10 shows a comparison of depth maps obtained from the frames depicted in Figure 3.4 using ToF only, Semi-global Matching (SGM) [77] stereo with rank filtering [78], and the local/TV stereo fusion. The local method was parametrized with $\alpha = (0.7, 1.5, 0.1)$ and $\gamma = (17, 0.1, 3)$, whereas $\mu = (5, 1)$ was used for the global method.

Besides the low resolution a considerable amount of noise can be observed in the ToF image, especially in the dark regions of the poster (A) and the foam plate (B). SGM fails on the wooden plates (C) due to lack of texture. Bleeding of disparities between the two statues (D) is also observable. Due to the fine texture on the poster (E), SGM estimates the right disparity in that region.

In general, the variational approach produces smoother results than the local one. This is to be expected due to the global regularization employed. Both fusion methods eliminate most of the noise around the poster (F) by using the available texture from stereo. The silhouettes (G) are reconstructed more precisely than in either ToF or stereo reconstructions alone and fine are details retained (e.g. the pyramid (H)). Also, the corner between the plates (I) that was corrupted due to multi-path effects is reconstructed properly as stereo cues were present in the corner. Conversely, erroneous reconstructions caused by multi-path effects can be observed on the table top (J). Both stereo and ToF systems estimate the wrong depth here. Therefore, it is not possible to improve the result using data fusion, as it was proposed above.

3.7.2 Evaluation with Reference Data

In the style of the Cornell Box [65], we[8] created the HCI-Box dataset containing different geometrical objects (cf. Figure 3.11). The box has dimensions of $(1.0 \times 1.0 \times 0.5)$ m. Due to the presence of large untextured surfaces and scarcely available horizontal gradients, this scene contains only few stereo cues. It can therefore be interpreted as an extreme test-case for the fusion techniques (i.e. fusion techniques should not produce worse results than either of the subsystems alone).

A synthetic model of the box was created with an error less than 1 mm. The extrinsic camera parameters of all three cameras, with respect to the 3D model, were obtained by manually selecting 2D-to-3D correspondences. Reference depth maps were obtained by rendering this model into the left camera view. With few exceptions, the reprojection error is lower than one pixel. These exceptions occur at depth edges (cf. Figure 3.11 right panel) due to small errors in alignment (cf. Section 3.8

[8] The HCI-Box was joint work with Henrik Schaefer [157] and Stephan Meister [126].

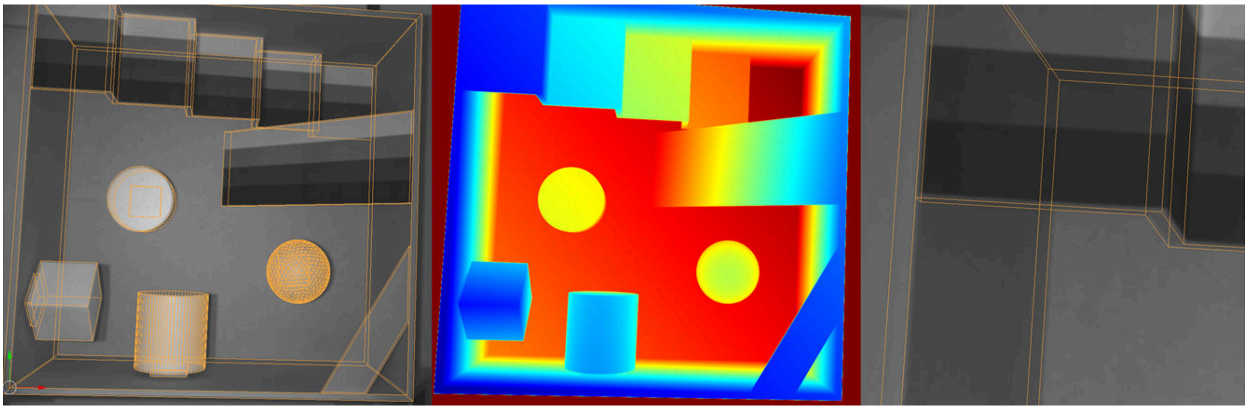


Figure 3.11: **Left**: HCI-Box reference target with overlayed reference mesh. **Middle**: Ground truth color-coded depth map. **Right:** Crop of area with largest misalignment between reference data and the stereo frame.

and Chapter 4). For quantitative evaluation, the disparity maps obtained by all methods were converted into metric depth maps. The variational approach was parameterized with $(\gamma_1, \gamma_2) = (5, 1)$ and the local method with $\alpha = (0.05, 0.05, 1.6)$ and $\gamma = (17, 0.1, 3)$. The difference between the reference and the obtained depth maps were then calculated (cf. Figure 3.12, right panel). Both methods show similar results with large errors on the box sides due to multi-path effects. For further evaluation, these regions of inter-reflection were masked out. The proposed fusion methods were compared with pure ToF upsampling and reprojection, SGM stereo as well as standard and adaptive TV regularized smoothing applied to the reprojected ToF data only, i.e. Equation (3.50) without the stereo data term [112]. It should be noted that ToF smoothing with adaptive TV does make use of edges obtained from the left stereo image, but not the right stereo image. The comparison was done by computing the absolute deviation of the computed depth maps from the reference depth map and then computing quartiles over the whole region of interest considered for evaluation. Note that these quartiles give information of the error in different parts of the scene and therefore do not necessarily correspond to the spread of per pixel errors. To assess the relative improvement of the

Table 3.1: Summary of GT evaluation on regions without multi-path. All values are in cm. Columns 3, 4 and 5 contain the quartiles of the absolute error distribution with respect to the ground truth. Small values are better. The last two columns contain the median per pixel improvement of reference data error between the method considered in the row with respect to local and global fusion.N.B. Large values are better.

Data	Method	1st Quart.	Median	3rd Quart.	M. i. loc.	M. i. glob.
TOF-data	upsampling	0.8	1.7	3.0	0.0	0.1
TOF-data	glob. meth./std. TV	0.8	1.7	3.1	0.1	0.1
TOF-data	glob. meth./adapt. TV	0.8	1.6	2.9	0.0	0.0
Stereo	SGM	0.8	1.8	3.2	0.1	0.3
Fusion	local method	0.8	1.7	3.0	0.0	0.1
Fusion	glob. meth./adapt. TV	0.8	1.6	2.8	-0.1	0.0

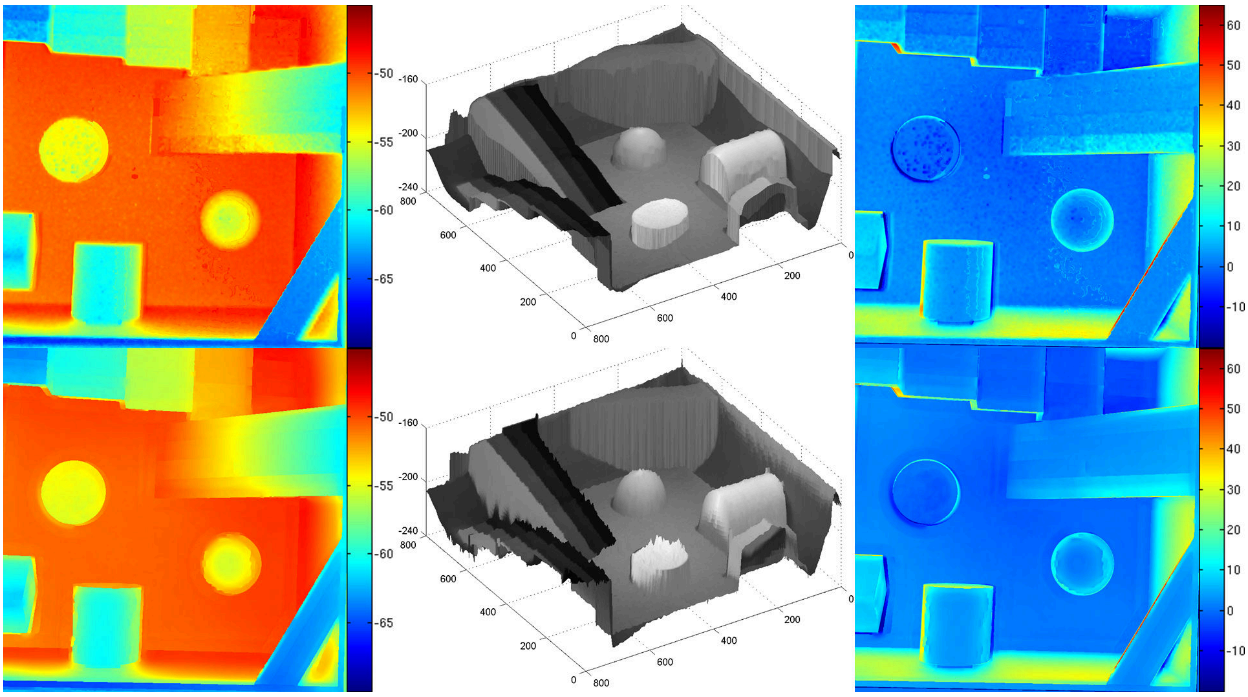


Figure 3.12: **Top:** Local method. **Bottom:** Variational method. **From left to right:** Disparities, 3D reconstruction and difference between reconstruction and GT in cm.

fusion techniques with respect to the baseline methods considered, the median (per pixel) decrease in absolute error between each method and the two proposed techniques were additionally considered. Results are summarized in Table 3.1.

As expected, the differences between the ToF only and the fusion techniques are rather subtle (improvement of $0 - 1$ mm) due to the sparsity of stereo cues to improve upon this result. Conversely, the improvement over pure stereo is slightly larger ($1 - 3\,\mathrm{mm}$). To give some further insight about the utility of the continuous variational method, relief plots (cf. Figure 3.13) along rows of the depth images were made comparing ground truth, variational fusion and SGM. The plots indicate that the variational fusion method produces results that are less corrupted and resemble the GT relief more closely than SGM. This can be seen on *a)* the stairs where the stereo results could be interpreted as a slope, *b)* the sphere and *c)* the slope. The negative effects of interreflection can be observed in *d)* and *e)* .

3.8 Summary and Outlook

3.8.1 Summary

This chapter dealt with data fusion of ToF cameras and passive stereo to overcome the limitations of the individual systems. The

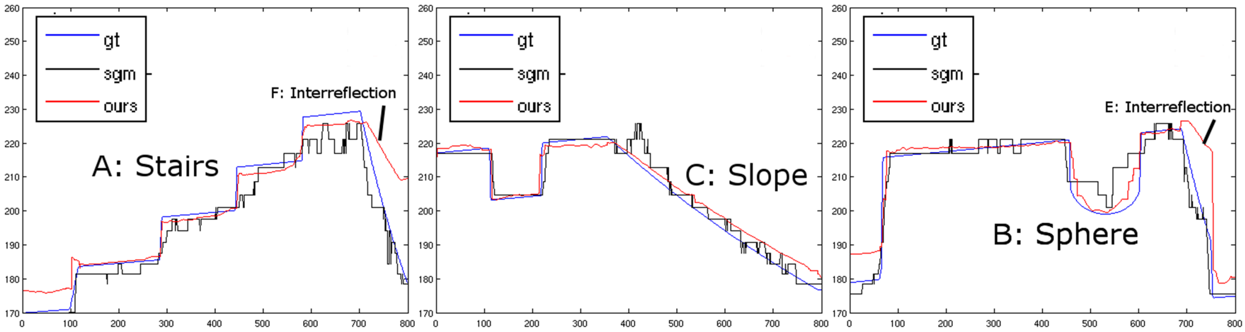


Figure 3.13: Reliefs of row 50, 280 and 500 depicting the stairs, the slope and the sphere. Note the artifacts created due to discretization and discrete regularization.

first part was a theoretical consideration of how ToF-stereo fusion techniques existing today relate to the measurement models of the individual subsystems. I presented the *symmetric early fusion* model and showed how existing *early* and *late* fusion techniques can be derived from this by quadratic approximation. To the best of my knowledge, this is the first derivation of this kind.

The second part of this chapter was devoted to the presentation of two *early fusion* techniques based on (heuristic) fidelity measures derived from the input images to guide the optimization process. The first method was a local algorithm based on block-matching, while the other one additionally made use of variational regularization. Both methods produce a high resolution depth map of the same resolution as the stereo system. Qualitative results showed that the resulting system displays many favorable properties such as robustness towards lack of image texture, robustness towards ToF noise if texture is present, accurate silhouettes and finally, the lack of occlusion artifacts. The main remaining issue that the fusion system cannot handle is multi-path (reflective surfaces). As these cause errors in both ToF imaging and stereo, one subsystem cannot compensate for the errors caused by the other one. This finding was confirmed during quantitative evaluation. To this end, a millimeter-accurate reference dataset [9] with little texture has been created. Results on this dataset indicate that the systems can robustly handle this extreme case by relying mostly on the ToF subsystem. A slight improvement could be observed in the global technique. This is mostly due to the adaptive regularization that makes use of the left stereo image as a similar effect was observed if the stereo data term was left out during inference.

[9] http://hci.iwr.uni-heidelberg.de/Benchmarks/document/hcibox/

3.8.2 Outlook

Being composed of many sub-modules, the development of the ToF-stereo fusion techniques naturally lead to many open questions, some of which motivated the following chapters and yet others that are subject to future work.

Understanding Alignment Errors The alignment of the ToF-stereo system is subject to errors using traditional calibration methods due to the low resolution of the ToF images and remaining systematic errors. Therefore, the question arises how misalignment can be quantified and whether it may be possible to resolve it. In Chapter 4, I seek answers to the first question, whereas I discuss ideas towards the second one in the Outlook of that chapter (cf. Section 4.8.2).

Multi-Path The most severe artifacts observed were those, due to reflective surfaces because these cause errors in both stereo and ToF imaging. This limits the usability of such systems if not understood and accounted for. A first step towards understanding multi-path effects by means of reproducing them in simulated data was undertaken in [125]. Reflections also motivate the work presented in Chapter 6, which is concerned with understanding and handling reflections in stereo. A question that still remains is how these insights can be used to create a practical fusion system.

Symmetric Full Model Inference The existence of the full symmetric fusion model was a rather late insight and naturally lead to the question, whether it is possible to directly solve this model, and if yes, whether solving such a model can in any way improve on current fusion techniques. There is evidence that this may be the case as it is easier to actually measure the correct model constants. Moreover, the noise characteristics of the ToF measurements are much simpler to model. Finally, it is naturally possible to resolve range ambiguity using the full model as the periodicity of the ToF measurements are retained (cf. Chapter 5).

Quantitative Evaluation While many ToF-stereo fusion techniques exist today, it is difficult to decide which one to choose for a certain application due to a lack of comparative studies. The leading questions here are:

- How should one choose the evaluation datasets, as we have seen that the choice of scene heavily influences the algorithm performance.
- What kind of performance metrics and experiments should be used for benchmarking?
- Where do we get reference implementations for the existing techniques?

There are many possible solutions and discussing all of them is out of scope of this chapter. However, I do discuss ideas and concepts in detail in [133] and [132] and refer to these works for anybody interested.

4 Uncertainty Estimation for Alignment of Stereo and Range Data

The following Chapter is based on my work previously published in [102].

4.1 Motivation

Many applications require (range) information from two different frames of reference (i.e. measurement systems) to be combined in some way nor the other. As an example, the majority of fusion techniques presented and reviewed in Chapter 3 reproject the ToF range data into the stereo frame to generate the initial depth map and prior. The application in the following chapter is the generation of reference data for performance analysis of stereo matching algorithms. Reference data is needed when quantitative performance evaluations are a requirement; this is for example the case for safety-relevant applications such as driver assistance systems. Here, range data is obtained by a measurement modality of higher accuracy such as LIDAR or structured light. This data is then projected into the stereo camera frames to obtain reference disparity maps, which can subsequently be compared with the output of a stereo algorithm.

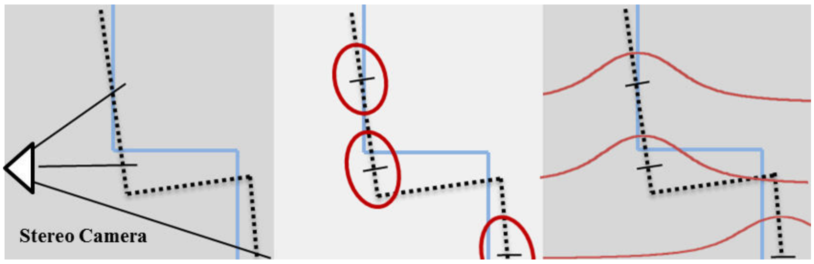


Figure 4.1: Schematic Illustration of Pose Estimation Error Effects. **Left**: The blue line is a schematic ground truth depth section along a pixel row, while the black dotted line represents misaligned range data from Range measurement device (LIDAR, ToF) projected into the stereo reference frame. Pixel positions of interest marked by solid black crosslines. **Middle:** The reprojected initial depth estimates have a localization error (red ellipses) in the stereo image space e.g. the depth assigned to the center stereo pixel actually belongs to one of the neighboring pixels. **Right**: Using the Range measurement error as the per pixel uncertainty without accounting for the localization errors underestimates pixel space uncertainties in some cases(middle line).

The relative pose between range and stereo frame required for reprojection as well as the range data itself are obtained from measurements. Measurements are subject to statistical measurement errors. This fact is well known and accepted in physics and the photogrammetry communities. I still stress that *every* measurement has an associated error or measurement uncertainty, as it is often overlooked or ignored in vision research. Hence, there also will *always* be an error or some level of uncertainty in the reprojected depth maps that are computed from the range data and estimated relative pose. It is important to quantify this uncertainty for both benchmarking and fusion purposes. In the latter, enforcing a ToF depth prior in areas with a large uncertainty in the reprojected ToF data may falsely suppress the correct disparity value of a pixel (cf. Figure 4.1, right image). Similarly, for benchmarking purposes, we have to understand that reference data is never perfect. It doesn't make much sense to compare the results of a stereo matching algorithm to reference data in areas where the reference data cannot be trusted. We need to understand the limits of the measurement devices in order to judge the quality of the reference dataset. Very often the uncertainty of the reprojected depth maps are specified as the range uncertainty of the measurement system (e.g. ToF, LIDAR) without taking pose estimation uncertainty into account. For reference measurements with LIDAR, the measurement uncertainties are an order of magnitude smaller than the typical error of a stereo system. This subsequently leads to statements which justify the omission of supplying uncertainty estimates with the reference data. Figure 4.1 illustrates that this can lead to incorrect estimates. When projected into a different camera frame, the errors in pose estimation result in an error in the localization of projected 3D points in the stereo frame. While the effect is not large in homogeneous areas, it causes large misalignment errors at depth discontinuities that can well be larger

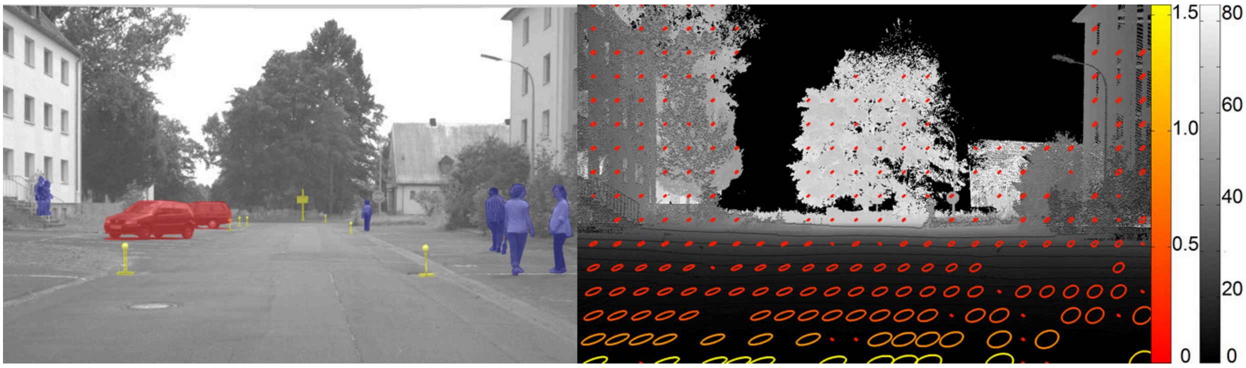


Figure 4.2: Reference Data Needs Error Bars. **Left:** Left stereo image with overlay of dynamic objects. **Right:** Ground truth disparities obtained by projecting LIDAR measurements into the left frame. The sparse overlay ellipses indicate the uncertainty in the localization of the projection in the image space. The error in the disparity is encoded in the color of the ellipse. Since the measured reference data is always subject to measurement errors the resulting ground truth dataset will also be subject to uncertainty.

than the uncertainty of the stereo system alone. Unlike existing methods proposed in the fusion literature, which rely on heuristics or learning to account for these effects, I present a rigorous treatment of error estimation and propagation (cf. Figure 4.2) to obtain meaningful per-pixel uncertainty distributions for reprojected depth maps.

4.2 Contributions

For data acquisition, a high-end camera stereo system was placed in a car[1]. The systems acquired sequences from an urban street scape. The same area was reconstructed using the best LIDAR mapping system available for this task (cf. Figure 4.4) by colleagues from the group of Prof. Claus Brenner at the IKG in Hannover[2]. My contribution to this project was the processing pipeline and error analysis of the data after acquisition. The aim was to focus on accuracy: How accurate can real-world ground truth become at individual pixels when all involved systems are state of the art? Although the approach I present, generalizes to arbitrary 3D scanners and camera setups in static scenes, the focus was on large-scale outdoor scenes ($> 30.000\,\mathrm{m}^2$) common in automotive applications. These can to date only be acquired by LIDAR mapping systems.

I present an approach to obtain ground truth reference data and per-pixel uncertainties thereof. The process is illustrated in Figure 4.3 and can be divided into the following steps: The static scene is scanned first and then a calibrated stereo sequence is recorded within this scene. The camera location for each frame

[1] The stereo acquisition was overseen by Stephan Meister and Wolfgang Mischler from our lab.

[2] `http://www.ikg.uni-hannover.de/index.php?id=764&L=gtizhuodyalitnaq` .

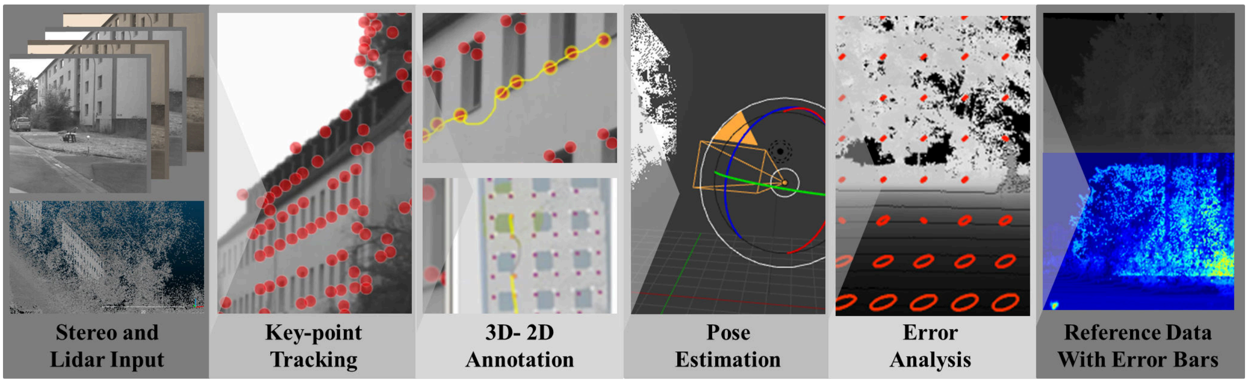


Figure 4.3: Workflow Stages: Starting with a LIDAR scan and an image sequence, we compute 2D feature tracks. These are matched with landmark 3D points using manual annotations (Section 4.4.1(.1)). Using these annotations and the other 2D feature tracks, we estimate the pose of each frame (Section 4.4.1(.4)). By means of covariance analysis and uncertainty propagation, we then obtain uncertainties in the localization of the reprojected 3D point cloud (Section 4.5). We then combine these localization uncertainties with the reprojections to finally output reference disparity maps and per pixel disparity distributions (Section 4.6).

is estimated locally with respect to the LIDAR frame, based on manually selected 2D-3D-correspondences. All cameras and correspondences are inserted into a bundle adjustment model, considering all error sources appropriately based on Gaussian errors in 2D feature localization, LIDAR accuracy and camera calibration parameters. Finally, the covariance of the bundle adjustment functional was evaluated at the solution to assess the uncertainty in the derived camera extrinsics. The resulting error distributions of the inputs (LIDAR, image data, intrinsics) and derived inputs (extrinsics) are propagated to obtain a localization error of LIDAR points in image space. Subsequently, these are converted and integrated over to obtain per-pixel uncertainty distributions of the reference disparity image. As a result, the method I propose comprises a full error propagation, starting with Gaussian error assumptions of the involved measurement devices and ending at per-pixel non-parametric disparity distributions. The subsequent pages are organized as follows: After presenting the related work in the next section, I describe the acquisition and processing pipeline in Section 4.4 with focus on my contributions to the project: annotations, data processing and how uncertainty can be estimated. Section 4.5 then describes the error propagation required to obtain the localization errors of 3D point projections. In Section 4.6, I describe how these localization errors can be converted into per-pixel uncertainty distributions for the disparity. Finally, before concluding the chapter with a summary and outlook of the next steps, I give some insight into how performance analysis of stereo data can benefit from reference data with error bars in Section 4.7.

4.3 Related Work

The related work is separated into three parts. First, I discuss the related work on techniques to generate reference data. Next, I expatiate on existing stereo datasets and the error estimates the creators provided wherever applicable. Finally, I present prior work on uncertainties for bundle adjustment problems that strongly influenced the work presented.

Generation Techniques: Ground truth generation implies two parts: an evaluation dataset and a reference dataset with superior accuracy. Different techniques differ in the way these datasets are obtained [101].

Synthetic imagery [137, 73, 30] allows for generation of reference data with little uncertainty and makes white box testing[3] of algorithms feasible by varying parameters such as geometry, light and materials. Yet, it remains to be shown whether content and renderer can model reality well enough [122, 69].

Another option is to record real data and use *manual annotations*. While relatively new to low-level vision, efforts have been undertaken with some success [114]. With the advent of crowdsourcing platforms [47], generation of such data has also become scalable. While the accuracy is reported to be good in general, possible biases introduced by humans are yet to be investigated. Finally, reference data can also be obtained by *measurement*, e.g. by using more than two cameras [128], additional devices such as the Kinect [123], a LIDAR scanner [60], or by using multiple exposures and UV-paint as in [10]. The approach of using more data from the same modality and reducing the data to create 'measurements' is not as costly as using dedicated measurement devices and sometimes scales very well because existing vision algorithms only need to be slightly modified. It should be noted, however, that in any case the reference data is itself obtained by measurement and therefore subject to uncertainty. Assessing this uncertainty is of utmost importance as statements such as "LIDAR is always more accurate than stereo" do not hold in general [173].

Stereo Datasets[4]: General-purpose real-world reference data has been published in the Middlebury database [10] with an estimated accuracy of around 1/60th of a pixel. This value is

[3] Measuring algorithm performance as a function of scene parameters such as weather/lighting conditions, number of people, etc. .

[4] Although most of the following works comprise additional datasets next to stereo data, I only focus on the latter.

derived from assumptions on the used block matching scheme and a down-sampling of originally larger images.

The EISATS database comprises a variety of sequences both real and synthetic [128, 181]. Using a third camera in the real dataset for additional redundancy proved to be beneficial for achieving an improved quality, but the accuracy of this data has not been thoroughly evaluated.

The closest approach to the one described here in terms of experimental setup is the one used for KITTI dataset [60]: Here, a stereo setup was combined with a car-mounted laser scanner. Mounting a LIDAR on the car has two main advantages. The scene can be recorded both in 2D and 3D at the same time and the density of 3D measurements is maximized as the LIDAR is very close to the optical axis of the stereo cameras. A disadvantage is that the system is moving while scanning, introducing a possibly low point density at high speed as well as motion artifacts. Although the accuracy was not explicitly evaluated in the original publication, it is reported by the authors to be less than three disparities for most of the pixels.

In our approach, the scene is scanned first. The stereo datasets are then acquired separately later. Hence, motion artifacts cannot occur and the sampling is roughly spatially uniform. In both KITTI and our setup, LIDAR was chosen as the most accurate and viable option to obtain depth in large scenes. Note, however, that our approach can be applied to any measurement technique with known uncertainty. Also the focus of all these databases is the creation of the ground truth database and the evaluation of algorithms. The work presented here aims at exemplifying error bar computation for real-world stereo ground truth using an appropriate statistical model.

Finally, the work most similar to the work presented in terms of scope is [173]. Here, uncertainties in camera intrinsics/extrinsics, LIDAR measurements and image key-point estimation are propagated to obtain reconstruction uncertainties for multiple-view stereo. While the authors make extensive use of sampling to estimate uncertainty, we provide an analytical solution for both camera pose estimation and the uncertainty of the disparity maps. For the first time, this allows for handling large numbers of frames (more than 1000 vs 25 in [173]). A comparison between a reimplemented version of their method with the proposed method shows a considerable speed up, even for small

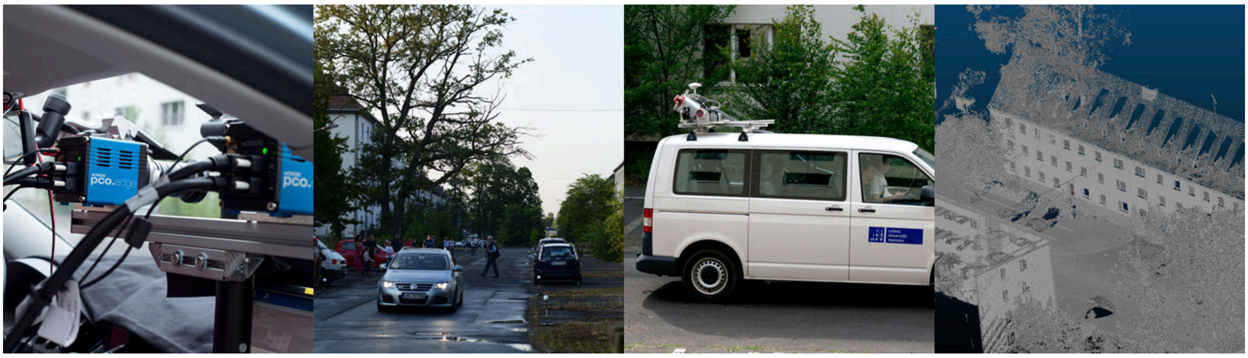


Figure 4.4: Experimental Setup.**From left to right:** Stereo rig, set photo, LIDAR mounted on car and resulting data. A video showcasing the collected data canbe found at `http://hci.iwr.uni-heidelberg.de/Benchmarks/document/StereoErrorBars`

problems. Moreover, using the proposed method yields tighter bounds on the camera pose uncertainty (cf. Section 4.4.2).

Uncertainty Estimation for Bundle Adjustment: A rich body of work exists on the theory of uncertainty estimation in the related field of bundle adjustment [96, 95, 179, 55]. Most techniques use local features of the bundle adjustment energy in the optimum, e.g. covariance analysis. A lot of effort is then put into tackling the inherent gauge ambiguity[179] issue of the structure from motion problem. The work presented here is inspired by these works. Yet, I am able to use a bundle adjustment variant for estimating the camera parameters that circumvents this gauge ambiguity by fixing the gauge to the LIDAR reference frame. Also it should be noted that the final goal is not the reconstruction of the camera parameters, but rather the generation of stereo disparity maps with a per-pixel uncertainty. To assess the quality of our camera reconstructions I relied on work from [45].

4.4 Ground Truth Acquisition

The acquisition modalities are depicted in Figure 4.4. A reference 3D point cloud of a street scape was collected using a RIEGL VMX-250-CS6 mobile mapping system. The stereo system consisted of two cameras with a 30 cm baseline equipped with 12 mm lenses. With a sensor size of 16.64 mm×14.04 mm, this corresponds to a field of view of 69.5°. The image sequences were acquired at 200 Hz with a resolution of 2560×1080 pixels. Preprocessing steps of the stereo data involved a lossless compression [85] of the 16 bit pixel data to 8 bits as well as camera calibration using [1]. Further details of the acquisition system can be found in the supplemental material of [102].

4.4.1 2D-3D Alignment

All measurement based reference data acquisition systems rely on a 2D-3D alignment step at some point of the processing pipeline. I will review the basic pose estimation and calibration process to build on this step for both explaining our alignment process as well as on how we derive error bars.

With K, I refer to the set of possible internal camera parameters and with $\mathbf{so}(3)$ to the group of rotations. For a distortion-free perspective camera with four parameters[5], $K = \mathbb{R}^4$. Let

$$\pi : (\mathbf{X}, \mathbf{t}, \kappa) \to \mathbf{x}, \tag{4.1}$$

$$\mathbf{X} \in \mathbb{R}^3, \mathbf{t} \in \mathbf{so}(3) \times \mathbb{R}^3, \kappa \in K, \tag{4.2}$$

be the projective mapping of point $\mathbf{X}$ from the world to image coordinate system using the extrinsic parameters $\mathbf{t}$ and intrinsics κ. Furthermore, let $\{(\mathbf{X_i}, \mathbf{x_i^j})\}$ be a set of 3D-2D correspondences of p measured 3D points $\mathbf{X_i}$ and their projections $\mathbf{x_i^j}$ in the jth frame of an image sequence containing n images. Then, the optimal intrinsic parameter κ^* and set of extrinsics $T^* = \{\mathbf{t^j}^*\}$ for each of the n frames is given by

$$(T^*, \kappa^*) = \underset{T,\kappa}{\operatorname{argmin}} \sum_{j=0}^{n} \sum_{i \in V(j)} \left\| \pi\left(\mathbf{X_i}, \mathbf{t^j}, \kappa\right) - \mathbf{x_i^j} \right\|^2, \tag{4.3}$$

where $V(j) \subset [0, ..., p]$ is the subset of 3D points that are visible in the jth frame. For a fixed camera - LIDAR setup such as KITTI this is done once in a calibration step with calibration targets before acquisition. Both geometry and projection of salient points are known here such that P can be obtained automatically. In our case, the LIDAR and the camera rig measure independently. This has the advantage of having LIDAR data at a much higher point density. In addition, it allows for capturing image sequences from other camera modalities (e.g Time-of-Flight, Plenoptic cameras) without requiring all cameras to be mounted on the same rig. In this setup, however, 2D-3D correspondences cannot be automatically aligned beforehand anymore. Picking individual landmark points out dense projections of point clouds (i.e. using a point cloud viewer)is an extremely tedious and error-prone task, as projections of points very far from each other can be in close proximity in screen space.

Parameter	Value
Detector Type	Harris Corners
Matching Type	Cross correlation
Matching window	21×21
Search Neighborhood	21×21

Table 4.1: Voodoo Tracker Parameters

[5] horizontal vertical focal lengths (f_x, f_y) and principle point (c_x, c_y)

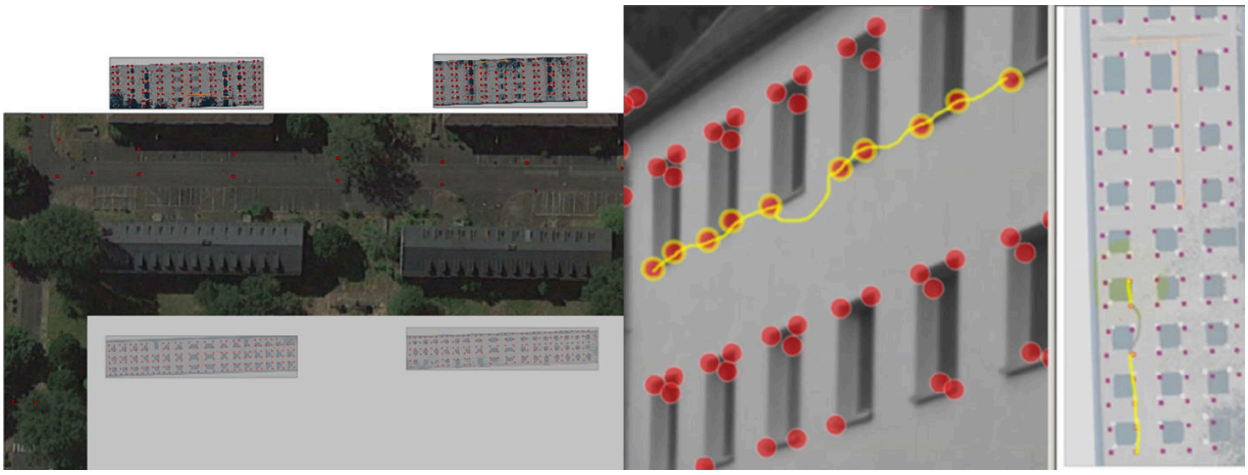


Figure 4.5: Map Annotation. **Left**: 3D landmark points obtained from the point cloud displayed as fold-out map. **Right**: Corresponding feature tracks being linked. A usage video can be found on `http://hci.iwr.uni-heidelberg.de/Benchmarks/document/StereoErrorBars` at 1:25min.

I propose an annotation and processing pipeline minimizing the risk of false correspondences (cf. Figure 4.3)

2D-3D Correspondence estimation/annotation

2D feature tracks $(\mathbf{x}_i^j)$ were automatically obtained with Voodoo Tracker[6] using the Harris Corner detector and a cross correlation based feature tracking (cf. Table 4.1). A subset of the tracks was matched manually with 3D points. This is difficult since each point in the 2D projection of the cloud corresponds to many 3D points at different depths. One solution would be to automatically mesh the point clouds, but it turns out that current approaches do not work well enough on our kind of data and also modify the location of the points in a non-linear way introducing unknown biases to the measurements. To ease point-picking, the 3D point cloud was reduced to a 2D representation in two steps:

Map Annotation For basic 2D to 3D registration, landmark 3D points from the LIDAR data are manually linked to corresponding key point tracks in the 2D dataset. This information is then later used to fix the gauge in the bundle adjustment problem. Currently, commercially available solutions[7] require a direct picking of points from a 3D point cloud. In practice, we found that doing so requires an experienced operator and is tedious as well as error-prone since it is difficult to pick 3D points from 2D projections thereof. This is especially true considering the amount of data we plan to handle. Therefore, we

[6] http://www.digilab.uni-hannover.de/docs/manual.html

[7] cf. http://www.thepixelfarm.co.uk/products/PFTrack

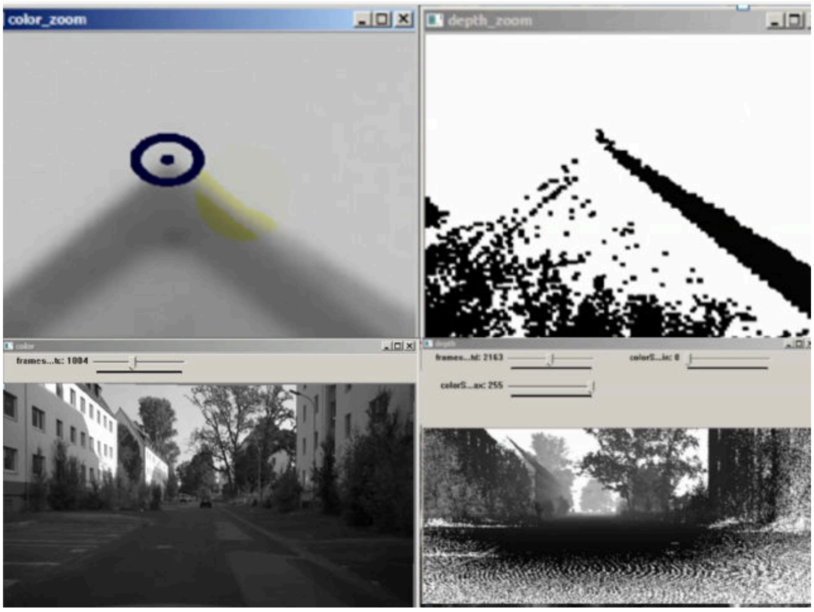


Figure 4.6: Range Annotation. Once an initial pose estimate is obtained additional correspondences can be made between the range image and the image sequence.

first picked corners of windows as well as markers placed in the scene using CloudCompare[8] once in the beginning of the whole process. These points can be tracked in the images quite easily. The picked points were placed in a fold-out map of the scene (cf. Figure 4.5). In our experience, displaying the 3D point cloud this way considerably simplifies and speeds up the annotation process.

Range Annotation Once enough points have been annotated, a rough pose estimation step is undertaken using the `solvePNP`[9] functionality in OpenCV [25]. The camera extrinsics obtained here are used to render a range image, in which each pixel corresponds to a maximum of one 3D point in the point cloud (cf. Figure 4.6). This range image is used to find additional correspondences in areas where not enough land mark points were found and on the other hand, to further refine the initial guess used in the bundle adjustment problem.

Camera Estimation With Known Variances

Neither the feature tracks nor the 3D points or internal camera parameters are perfect. Also the intrinsic calibration routine usually delivers a good initial guess $\hat{\kappa}$ for the intrinsics. I assume

[8] http://www.danielgm.net/cc/
[9] A method to solve the extrinsic calibration problem presented in (2.52)

Gaussian errors in each of these values:

$$\begin{aligned}\mathbf{X_i} =&\quad \mathbf{Z_i} + \epsilon_{\mathbf{X_i}} &&, \epsilon_{\mathbf{X_i}} \sim \mathcal{N}_3(0, \Sigma_{\mathbf{X_i}}), && (4.4)\\ \hat{\kappa} =&\quad \kappa + \epsilon_{\kappa} &&, \epsilon_{\kappa} \sim \mathcal{N}_4(0, \Sigma_{\kappa}), && (4.5)\\ \mathbf{x_i^j} =&\quad \mathbf{z_i^j} + \epsilon_{\mathbf{z_i^j}} &&, \epsilon_{\mathbf{x_i^j}} \sim \mathcal{N}_2(0, \Sigma_{\mathbf{x_i^j}}). && (4.6)\end{aligned}$$

To accommodate for these errors we modify Equation (4.3):

$$(\{\mathbf{Z_i}\}^*, T^*, \kappa^*) = \underset{(\{\mathbf{Z_i}\}, T, \kappa)}{\operatorname{argmin}} \; \Phi(\{\mathbf{Z_i}\}, T, \kappa), \tag{4.7}$$

with

$$\begin{aligned}\Phi(\{\mathbf{Z_i}\}, T, \kappa) = \sum_{j=0}^{n} \sum_{i \in V(j)} \Bigg(& \quad \left\| \pi\left(\mathbf{Z_i}, \mathbf{t^j}, \kappa\right) - \mathbf{x_i^j} \right\|^2_{\Sigma_{\mathbf{x_i^j}}} \\ & + \quad \|\mathbf{X_i} - \mathbf{Z_i}\|^2_{\Sigma_{\mathbf{X_i}}} \\ & + \quad \|\hat{\kappa} - \kappa\|^2_{\Sigma_{\kappa}} \Bigg).\end{aligned} \tag{4.8}$$

Here, $\|\mathbf{a}\|^2_{\Sigma}$ denotes the squared Mahalanobis distance

$$\|\mathbf{a}\|^2_{\Sigma} = \mathbf{a}^T \Sigma^{-1} \mathbf{a}, \tag{4.9}$$

with inverse covariance matrix Σ^{-1}. Note the quadratic penalty terms in Equation (4.8) and explicit usage of latent variables $\mathbf{Z_i}$ and κ. These are required as the first residual term is not linear in $\mathbf{X_i}$ and $\hat{\kappa}$, whereas it is in $\mathbf{x_i^j}$. This splitting of variables is often used to enable a better treatment of nonlinearities in Gaussian energy functionals [3, 190]. Also note that the first term corresponds to a bundle adjustment problem and the last two terms to priors on $\mathbf{X_i}$ and $\mathbf{x_i^j}$. In the optimization, it is therefore possible to include 2D feature tracks without 3D correspondences. Parameter estimation was done using the open source Ceres Solver [4] library.

4.4.2 Consistency And Precision of the Pose Estimation With Synthetic Data

Solving 4.8 together with the covariance estimation done in Section 4.5 yields a pose estimate together with an uncertainty estimate thereof. To assess the precision and consistency of the pose estimation system, I borrow ideas from [45]. Here, a method is proposed to compute consistency and precision of a dataset with respect to a reference dataset with lower but non-zero uncertainty. As the output of our system has the highest available precision, I had to resort to synthetic data and make some changes to the formulas in [45] to cater to the zero uncertainty of our reference.

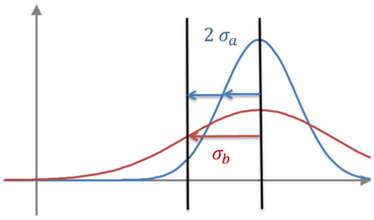


Figure 4.7: Illustration of Consistency. Both methods (blue and red) estimate the same mean and have the same absolute distance from the ground truth distance (left line). The blue method reports a smaller uncertainty for the estimate as compared to the red one. Yet the red method is more consistent with the reference value as the likelihood of the reference is larger (i.e. the Mahalanobis distance is smaller).

Consistency is a measure for the likelihood that both reference and synthetic datasets have the same parameters. As in [45], we report the Mahalanobis distance (cf. (4.9)) between the synthetic reference and the methods using the estimated pose covariance (cf. Figure 4.7), as described in Section 4.5.

Precision refers to the certainty of the method in the correctness of its parameter estimate. Given two parameter estimates with a similar consistency with regard to the reference, the estimate with the smaller uncertainty should be favored. Here, I report the trace of the estimated covariances (cf. Figure 4.8).

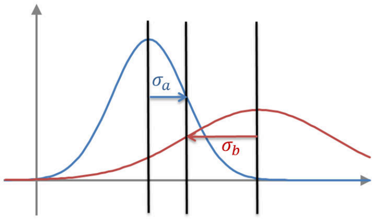


Figure 4.8: Illustration of Precision. Red and Blue methods have the same Mahalanobis distance. In this case the blue method with the smaller standard deviation has a larger precision value and should be preferred.

Table 4.2 summarizes the results. The reference data was generated by randomly picking p key points in the first frame, randomly choosing a depth for each key point between 5 and 70 meters and finally, by rejecting 3D points not visible in the $n-1$ other camera frames. The evaluation dataset was obtained by adding Gaussian noise according to the *noise* column of key point position and 3D point. I compare the method presented to a sampling based strategy similar to [173]. Here the 2D and 3D points are perturbed around the estimated solution (s times). After that, the best new parameter set is obtained by minimizing the bundle adjustment functional (cf. Eq. (4.8)) (keeping 2D and 3D measurements fixed). The pose and pose uncertainty are then obtained by and estimating the sample mean and covariance over the s bundle adjustment solutions. In the result columns, the mean consistency, the precision and the run times in seconds after 30 runs are reported. The standard deviation over the 30 runs for consistency always was around 1 and for precision and run time an order of magnitude smaller than the reported values. While we observe mostly similar consistency values between both methods - with the sampling consistency deteriorating with higher noise levels and larger datasets -, the proposed method produces a tighter precision bound on the parameter estimate with much faster run times.

Noise [cm, px]	Number of points p	Number of frames n	Sampling s = 100	Sampling s = 1000	Ours
(5, 0.1)	100	5	(5.1, 3.9e-4, 0.4)	(5.1, 4.0e-4, 4.7)	(5.3, 1.1e-4, 0.1)
(5, 0.5)	100	10	(7.7, 1.7e-2, 0.8)	(7.6, 1.7e-2, 8.2)	(7.6, 5.5e-3, 0.2)
(1, 0.1)	1000	10	(8.1, 7.8e-5, 9.5)	(7.9, 7.9e-5, 96)	(8.5, 2.2e-5, 2.4)
(5, 0.5)	1000	10	(8.2, 1.8e-3, 9.5)	(8.0, 1.8e-3, 97)	(7.2, 5.2e-4, 2.1)
(0.05, 0.5)	20	200	(34, 1.2, 1.6)	-	(34, 8.8e-1,0.9)
(0.05, 0.5)	100	100	(25, 1.6e-1, 4.3)	-	(25, 4.5e-2, 3.3)
(0.05, 0.5)	200	100	(26, 9.4e-2, 9.2)	-	(24, 2.5e-2, 7.6)
(0.05, 0.5)	200	200	(35, 2.2e-1, 20)	-	(35, 4.9e-2, 23)

Table 4.2: Pose estimation results on synthetic data. The tuples reported in the right 3 columns correspond to consistency, precision and run time. Lower values are better. Fields marked with '-' were omitted due to prohibitive runtime.

4.5 Reference Data with Error Bars

Once the pose estimation in Equation (4.8) has been solved we can proceed in creating reference data by computing a range image based on κ, T and the LIDAR point cloud by means of Equation (4.1). This reference data contains holes with no information whenever no LIDAR measurements map to the corresponding pixel location. In the following, I consider the extended reference data mapping

$$\tilde{\pi}_b : (\mathbf{X}, \mathbf{t}, \kappa) \to (\mathbf{x}, d), \tag{4.10}$$

which not only computes the projected image location of a 3D point but also the disparity of this point given stereo baseline b. With $\mathbf{d} = (\mathbf{x}, d)$ I will denote the vector containing image coordinates and disparity. The subscript b is omitted in the further discussion as it remains constant for each sequence.

The inputs in $\tilde{\pi}(\mathbf{X_i}, \mathbf{t^j}, \kappa)$ are either measurements or values derived from measurements. As measurements always contain errors, the reference point $\tilde{\pi}(...)$ will also have an error. To assess theses errors quantitatively, error estimates for $\mathbf{X_i}, \mathbf{t^j}$ and κ need to be obtained first.

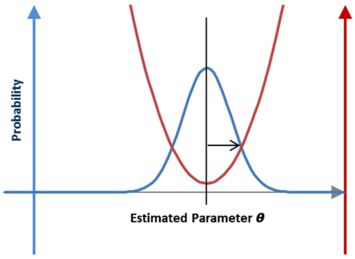


Figure 4.9: Illustration of Covariance Analysis. The least squares residual energy (red) corresponds to the negative log-likelihood of the posterior parameter probability (blue). The opening of the energy parabola obtained by Eq. (4.11) (black arrow) corresponds to the variance of the parameter estimate.

1. For the **3D point position** $\mathbf{X_i}$, I assume that the components are independently distributed such that $\Sigma_{\mathbf{X_i}} = \sigma^2_{\mathbf{X_i}} I$. In our case, this is the measurement error of the LIDAR scanner. For point clouds consisting of multiple LIDAR scans that were merged [60] via iterative closest points (ICP) or similar methods, the error should be the error propagated from the ICP fit.

2. For the **camera pose** $\mathbf{t^j}^*$, I assume that $\mathbf{t^j} \sim \mathcal{N}_6(\mathbf{t^j}^*, \Sigma_{\mathbf{t^j}})$. As $\mathbf{t^j}$ is a value derived from a least squares fit, $\Sigma_{\mathbf{t^j}}$ can be

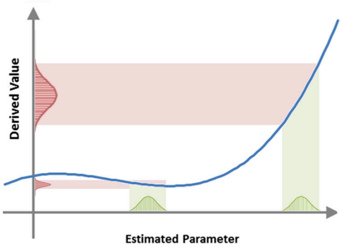


Figure 4.10: Illustration of Error Propagation. If a derived quantity (vertical axis, e.g reference point location) is a function of a parameter (horizontal axis, e.g. pose) which is subject to measurement errors (green distributions), the derived quantity will also have an uncertainty (red distributions) which is transformed according to the mapping between derived and original quantities. Linear error propagation can be applied if the function can be approximated linearly over 'most' of the input distribution (green and red overlay).

obtained by evaluating the covariance matrix of Φ at the solution $s^* = \{\mathbf{t^j}, ..., \}$ with

$$COV_\Phi(s^*) = (J_{\Phi(s*)} J^T_{\Phi(s*)})^{-1}. \tag{4.11}$$

Here, $J_{\Phi(s*)}$ is the Jacobian of the residual vector of Φ evaluated at solution s^* (cf. Figure 4.9). $\Sigma_{\mathbf{t^j}}$ is the diagonal block of $COV_\Phi(s^*)$ corresponding to the parameter block belonging to $\mathbf{t^j}$. Note that a regular bundle adjustment scenario has an inherent scale ambiguity which leads to $J_{\Phi(s*)}$ being rank deficient. In contrast, the functional presented has full rank as the scale is given by the 2D - 3D correspondences. Also note that by supplying the correct error estimates during the alignment fit, COV_Φ is properly scaled.

3. For the **camera intrinsics** κ, I either use the same approach as chosen for $\mathbf{t^j}$ or use variances estimated by external calibration tools. Again the distribution is assumed to be Gaussian with $\kappa \sim \mathcal{N}_4(\kappa, \Sigma_\kappa)$.

The error distribution in $\tilde{\pi}$ of the reference point and the error in the disparity measure can be obtained via error propagation. This is achieved either via sampling input realizations from the above distributions or by analytical linear error propagation (cf. Figure 4.10). For the latter, the full covariance matrix of the inputs evaluates to

$$COV_{IN} = \begin{pmatrix} \Sigma_{\mathbf{X_i}} & & \\ & \Sigma_{\mathbf{t^j}} & \\ & & \Sigma_\kappa \end{pmatrix}. \tag{4.12}$$

The error in $\tilde{\pi}$ is then obtained by linearizing $\tilde{\pi}$ at the reference point. Under assumption of a Gaussian distribution of the input variables the output is again Gaussian with covariance given by

$$COV_{\mathbf{d}} = J_{\tilde{\pi}(\mathbf{x},d))} COV_{IN} J^T_{\tilde{\pi}(\mathbf{x},d)}. \tag{4.13}$$

The choice between sampling and linear propagation depends on the available computational resources as sampling will deliver more accurate output error distributions given enough samples whereas linear error propagation is analytical and thus fast.

	Value	Uncertainty
INPUT	LIDAR (propagation)	$\sigma_{\mathbf{X_i}} = 1\,\text{cm}$
	LIDAR (pose uncertainty)	$\sigma'_{\mathbf{X_i}} = 3.5\,\text{cm}$
	Feature Track	$\sigma_{xij} = 0.5\,\text{px}$
	Focal Length	$\sigma_{K(f_x,f_y)} = 1.97\,\text{px}$
	Principal Point	$\sigma_{K(c_x,c_y)} = 1.46\,\text{px}$
OUTPUT	Pose (rotation)	$(r_x, r_y, r_z) = (3, 3, 2) \times 10^{-4}$ upper bound 0.026°
	Pose (translation)	(t_x, t_y, t_z) $= (1.23, 2.53, 2.17)\,\text{cm}$

Table 4.3: Summary of Input and Output Uncertainties used and estimated

4.5.1 Reference Data Sensitivity

In the following, an analysis of the reference data uncertainties will be given using the tools provided above. The values are summarized in Table 4.3. I start off by discussing the error values used for the inputs. The **LIDAR uncertainty** used for error propagation is obtained from the data sheet. For the contribution of the 3D points towards pose uncertainty (cf. Eq. (4.11)), a larger error has to be assumed due to the point spacing. Therefore, the localization of a manually picked point (e.g. a window corner) is only accurate up to the mean distance between points. This was determined by estimating the point density on building facades where the landmark points were chosen from. The **feature track accuracy** was empirically estimated, while errors in **focal length and principal point** were obtained from our calibration routine. For the **pose estimation accuracy**, I report the mean square root of the diagonal entries of $\Sigma_{\mathbf{tj}}$ obtained from covariance analysis for the translation over **100 frames**. The rotation is parametrized using a 3D angle-axis representation (cf. Figure 2.6) . The error has an upper bound of 0.026° based on the maximum deviation of the angle-axis vector. For a LIDAR point at 50 m distance, this corresponds to a localization error of around 2 cm. The error in the translation also amounts to 2 cm. Using the errors obtained from the input, the uncertainty in the reference data can be computed by means of error propagation. For each reference point, the full covariance in **d** (i.e. pixel localization and disparity error) was computed using both linear error propagation and sampling. In Figure 4.11, the square roots of the diagonal entries are reported for an example scene. The first two rows correspond to the localization error and the third row is the disparity error in logarithmic

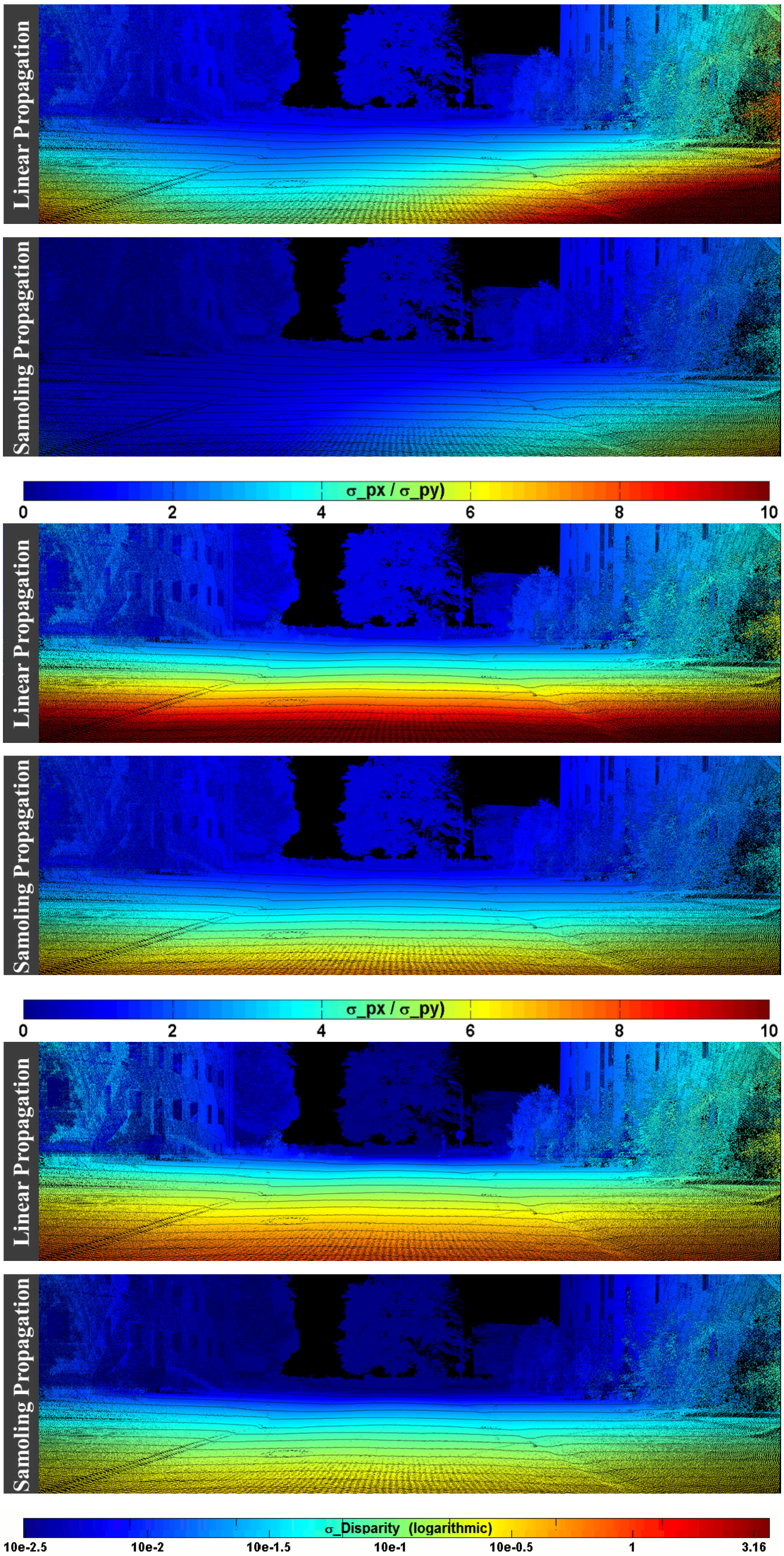


Figure 4.11: Diagonal entries of uncertainty $\Sigma_\mathbf{d}$ obtained by **linear error propagation** and **sampling**. **From top to bottom:** Localization error in x and y as well as disparity error of reference data points. Note that the bottom row is scaled logarithmically. While the general form of the error distribution is the same for both analytic and sampling based propagation, we obtain tighter bounds on all errors using sampling.

scale. For both linear propagation and sampling, we see the expected inverse distance reduction of all errors. While the disparity error for most parts is under one pixel, the localization error exceeds five pixels for points closer than a few meters. Also noticeable is the increase in x localization error towards the image edges observable in all our sequences. I believe that this is related to a rotational error of the camera localization. Finally, by comparing sampling and linear propagation, we can see that the sampling propagation in general gives a tighter bound on the reference data error while preserving the general shape. As both propagation methods yield similar results we conclude that linear error propagation can be used to obtain a quick though looser bound on the reference data error.

4.6 Disparity Maps with Error Bars

So far, I have discussed the reference data quality in terms of the localization and disparity error of each reference point. For evaluating a stereo algorithm, we are faced with a slightly different question as we are concerned with the question how good a given disparity map is. Hence, a distribution of possible disparity values in each pixel is required. Given a set of reference data points with uncertainty $R = \{(\mu_{\mathbf{r}}, \Sigma_r)\}$ computed as described in Section 4.5, I define the probability of a disparity map $\mathbf{D}$ to be

$$p(\mathbf{D}|R) = \prod_{\mathbf{x_i} \in \mathbf{D}} \frac{1}{N} \sum_{(\mu_{\mathbf{r}}, \Sigma_r) \in R} \exp\Big((\mathbf{x_i} - \mu_{\mathbf{r}})^T \Sigma_r^{-1} (\mathbf{x_i} - \mu_{\mathbf{r}}) \Big) , \tag{4.14}$$

with $\mathbf{x_i} = (\mathbf{p_i}, d)$ the disparity d at pixel position $\mathbf{p_i}$ and normalization N. The Gaussian distribution in Equation (4.14) is multivariate (in pixel position and disparity). This distribution can alternatively be computed by either sampling from the reference data distribution or analytically from the input data distribution directly using Gaussian error propagation. The main drawback of a linear error propagation is that the projection of Gaussian disparity distribution into image space yields multi-modal per-pixel distributions which cannot be accounted for using linear propagation. Figure 4.12 shows such distributions at example pixel locations. We can distinguish three error cases: First, due to extrinsic camera parameter uncertainty the locations of depth edges are projected to different pixel loca-

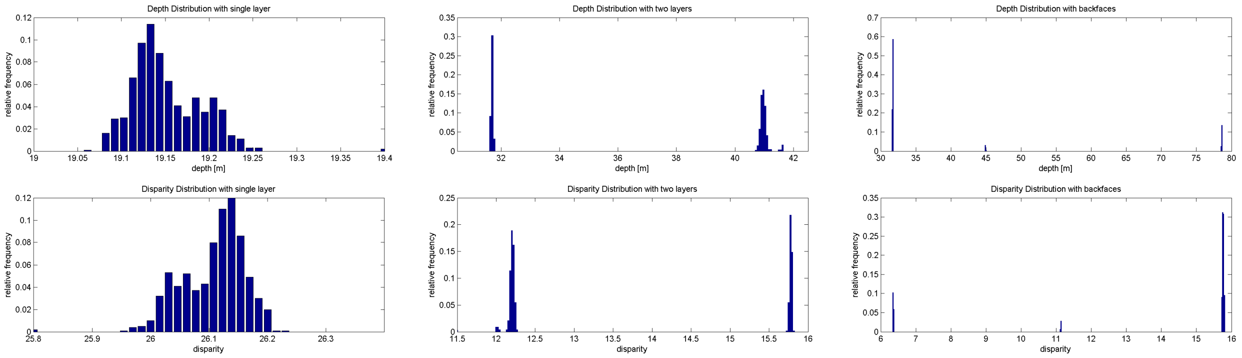


Figure 4.12: Example distributions on sampled depth maps (1000 samples). **From left to right**: pixel with single depth layer, edge pixel with two depth layers, pixel with unresolved back faces. **Top row**: depth distribution. **Bottom row**: disparity distribution.

tions. This causes bimodal disparity distributions since either the background or the foreground is sampled. The result is a very high variance, i.e. a large, though correct error bar on the ground truth.

Second, multi-modal distributions can occur caused by back-srufaces: multiple surfaces such as the front and back of a house as well as the houses in the background of the LIDAR point cloud are projected to the same pixel. This is a fundamental limitation of point clouds - yet, established meshing tools can not deal with our data as was explained in Section 4.4.1. In these situations, the ground truth is not wrong per se - but more reasoning is required to decide whether the multi-modality of the distribution is caused either by a depth edge or by back-srufaces.

Third, if the scanner did not measure a foreground object, for example due to limited resolution (e.g. landlines, small twigs on trees), the disparity distribution becomes unimodal but still displays the wrong depth of the object behind the small foreground object. This case can only be dealt with by more accurate measurement devices which do not yet exist for our application. The problem can only be alleviated by manual segmentation of foreground objects which are visible in the image, but not in the 3D scan.

Once the per-pixel distributions in disparity space are sampled, we can reduce their information to per-pixel scalar values. Figure 4.13 displays two such options: the top image contains the median of the disparity distribution. Assuming that the number of foreground samples outweighs the number of back-surfaces by a factor of at least two, this is a robust ground truth depth. Note however that this approach fails at depth boundaries when foreground and background can easily become equally likely. Therefore, the lower image displays the inter-

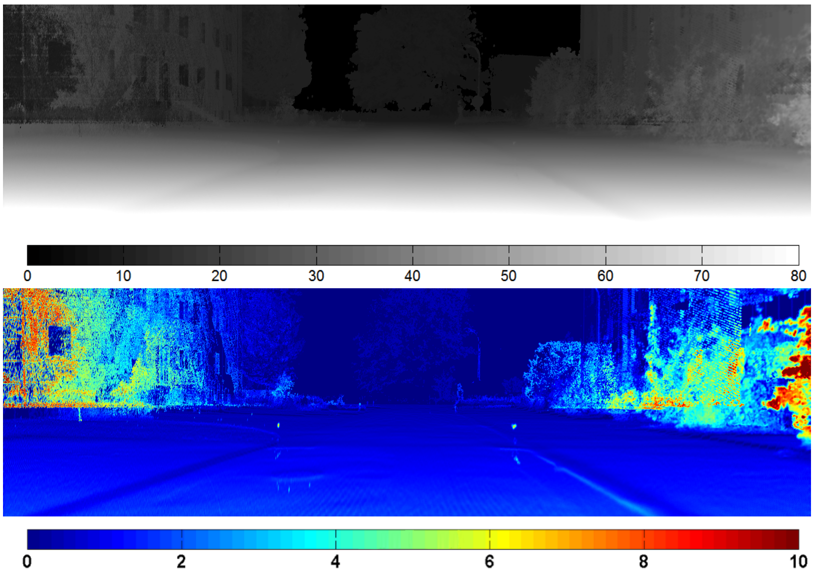


Figure 4.13: **Top**: Median of disparity distribution. **Bottom**: inter-quartile range of disparity distribution. High values show regions with unreliable ground truth mainly caused by vegetation and camera misalignments. Regions looking like artifacts are caused by backsurfaces as explained in the text. In all other regions, the inter-quartile range is below two disparities.

quartile range of the disparity distributions.

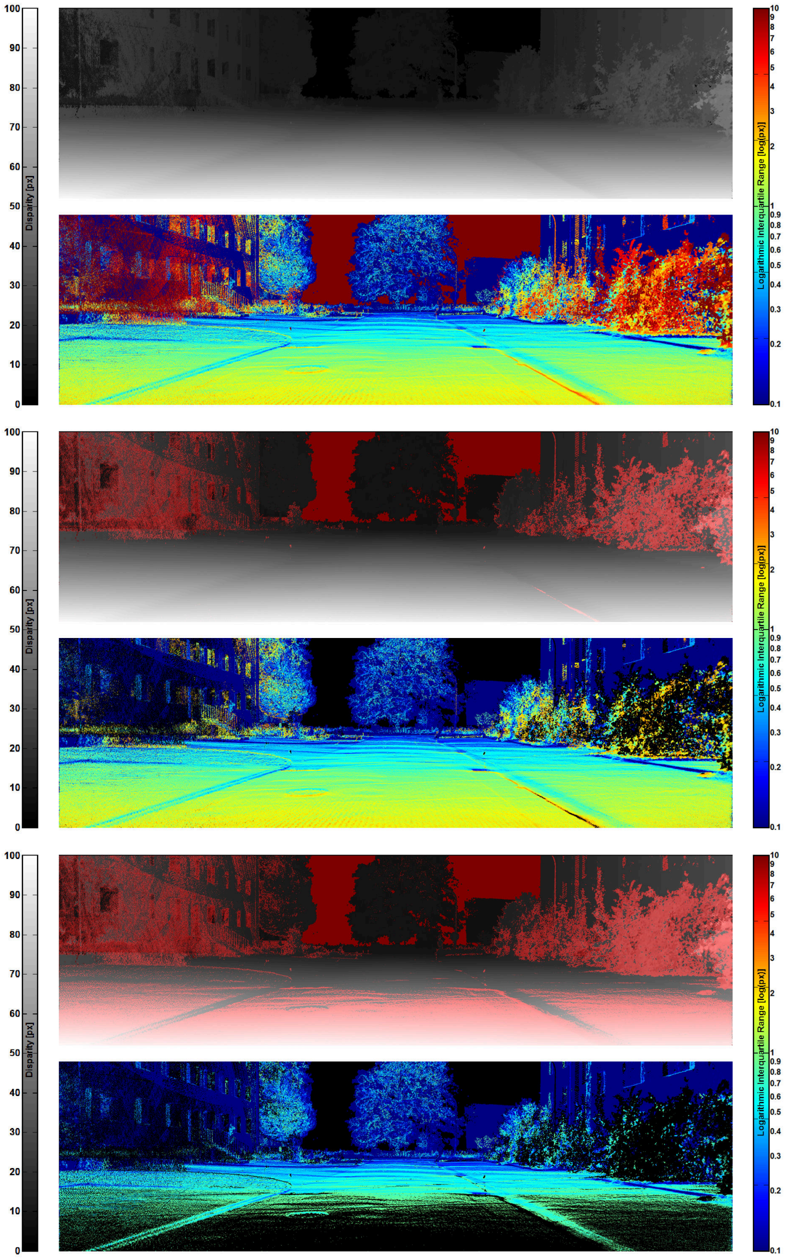
4.7 How can we use these Error Bars?

With the methods described in the previous sections it is possible to obtain reference data with per-pixel uncertainty distributions and to reduce the information to 'error bars'. A question that remains is how these error bars can actually further the field of stereo matching and performance analysis thereof. In the following, I will thus address this question. The way the output $\mathbf{D_s}$ of a stereo algorithm is usually benchmarked with reference data $\mathbf{D_r}$ nowadays is by computing the residual image

$$\mathbf{R} = |\mathbf{D_s} - \mathbf{D_r}|. \tag{4.15}$$

Then, some kind of pixel statistic of this residual image is computed and used as a scalar performance metric $\Phi(\mathbf{R})$. Metrics used include the mean absolute distance, the mean squared distance or the number of entries with a value larger than a fixed threshold over the whole image. For the latter, it only makes sense to include reference data pixels that themselves have an uncertainty lower than the threshold. Figure 4.14 illustrates the effect of applying different common thresholds to our data. It is important to mention that this type of masking is not necessarily the best option for performance evaluation. Since a smaller Φ implies a better algorithm, this invariably becomes the major op-

Figure 4.14: Sparsification of ground truth (top image of each pair) using measurement uncertainty (bottom image of each pair) with different thresholds. **Top pair**: no threshold. **Middle pair**: 3 px. **Bottom pair**: 1 px. Invalidated pixels are marked red in disparity image and black in uncertainty image



timization criterion. As long as the uncertainty of the reference data is smaller than the accuracy of the stereo algorithm this does indeed mean an advancement of the field. If the uncertainty is of the same magnitude or even larger, though, minimizing Φ only results in over-fitting to the reference. This effect has been observed in [60] where the authors note that “methods ranking high on Middlebury, perform particularly bad on [their] dataset” and “hope that [their] proposed benchmarks will complement others and help to reduce overfitting to datasets...”. Also, as it can be seen in our data, the uncertainty is not necessarily uniform over the whole image. In the following, we will show that using quantified per-pixel uncertainties can help identify such issues.

A simple performance metric based on the full distribution could be

$$m(\mathbf{D_s}|\mathbf{D}_r) = \sqrt{-\log(p(\mathbf{D_s}|\mathbf{D}_r))}. \tag{4.16}$$

Note that for Gaussian per-pixel distributions this term corresponds to the Mahalanobis distance between reference data and stereo output. For the following experiments, I consider the absolute residual difference (cf. Eq. (4.15)) and the uncertainty weighted absolute difference

$$\mathbf{C} = |\mathbf{D_R} - \mathbf{D_S}|/\mathbf{S_R}, \tag{4.17}$$

where S_R is a scalar uncertainty value such as standard deviation or interquartile range. Note the similarity between **C** and **R** and the consistency and precision values used in Section 4.5.

Experiments

For the experiments, the implementation of basic stereo algorithms provided by Scharstein and Szeliski was used[10]. Disparity maps on frame 4521 of sequence 0 were computed by various stereo algorithms. A summary of the results over all algorithms can be found in Figure 4.15. Following the remarks about the interpretation of the weighted image as a consistency value between ground truth and stereo algorithm, we can use the consistency and absolute difference images to gain further insights. Four different cases that are of particular interest. These are indicated as squares in Figure 4.15 and magnifications of the regions of interest described below can be found in Figures 4.17

[10] http://vision.middlebury.edu/stereo/code/

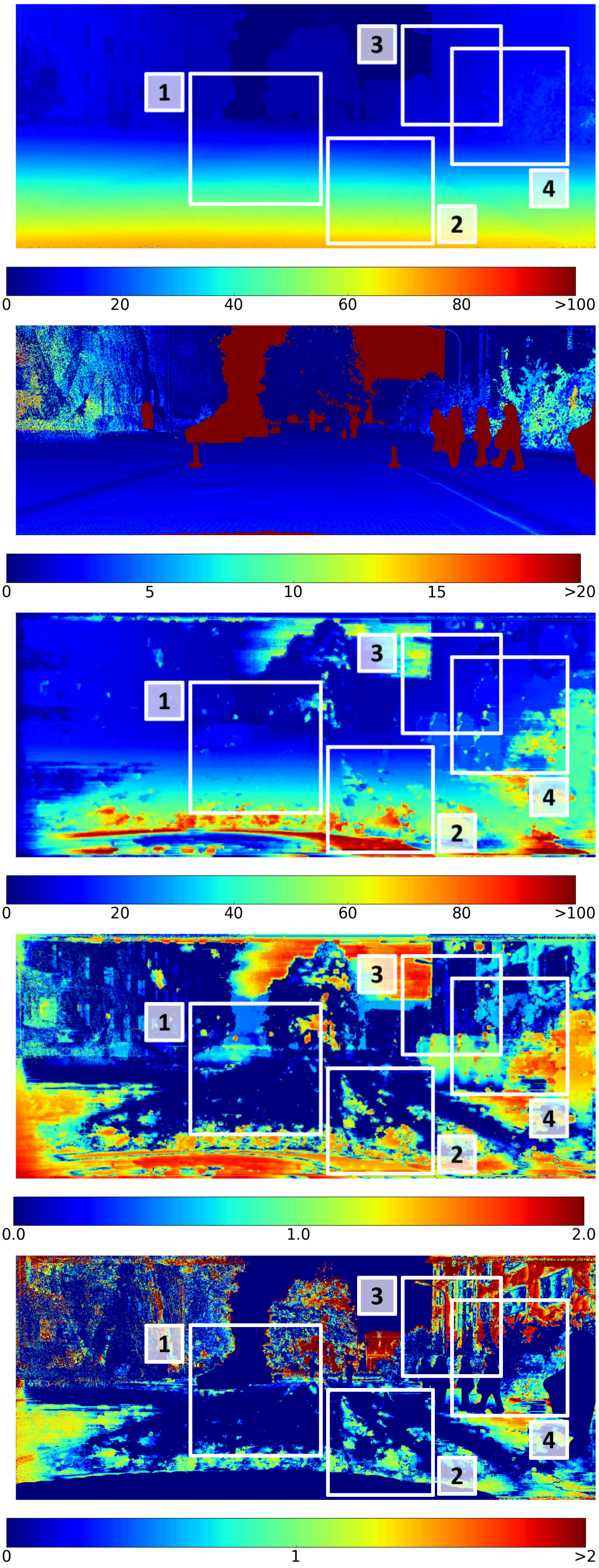


Figure 4.15: Frame 04521: **Top Row**: GT disparity. **Second Row**: Uncertainty map with dynamic areas masked out. **Third Row**: Average disparity over algorithms (DP, SP, SA, SGM, SSD, SSDmf). **Forth Row**: log_{10} of absolute disparity error **R**. **Bottom Row**: log_{10} of consistency **C**.. The regions of interest are marked by squares. regions 1 and 2 are discussed in Figure 4.16 and regions 3 and 4 in Figure 4.17

and 4.16.

1. The absolute difference is small and the consistency value large. This happens when there is a high agreement between stereo and ground truth. This can be observed on most of the street area (cf. Figure 4.16, left).
2. The absolute difference is large and the consistency value small. Here, we can be confident that the error is caused by the stereo algorithm. Such fail cases can also be observed on the street area (cf. Figure 4.16, right).
3. The absolute difference as well as the consistency value are small. While the stereo algorithm is close to the right result, we can again be confident that the small error is significant. In Figure 4.17 (left) this can be observed at the facade.
4. Finally, the absolute difference is large and the consistency value small. Here we can no longer trust our reference data[11] and should resort to other methods such as manual inspection. We observe this situation around the bushes in Figure 4.17 (right) where the LIDAR scanner delivered very noisy data.

Finally it should be noted that a more appropriate evaluation would require the stereo algorithm to propose a disparity distribution as well. Then, the performance metric would compare ground truth and computed disparity distribution, e.g. by a Kolmogorov-Smirnov test.

4.8 Summary and Outlook

4.8.1 Summary

I have presented a methodology to add error bars to image sequences with disparity ground truth. It is based on previously measured point clouds and arbitrary calibrated cameras and therefore highly versatile for all kinds of indoor as well as outdoor applications. However, due to the chosen 3D scanning device, the approach is limited to static scenes.

Based on intuitive inputs such as calibration, 2D feature and 3D LIDAR accuracy, I estimated the covariance matrix of our model at the solution to derive per-pixel depth-distributions. The results were used to define error bars, e.g. by computing

[11] Note, that this does not mean that the stereo algorithm is more accurate. We just cannot make any statements using the reference data here.

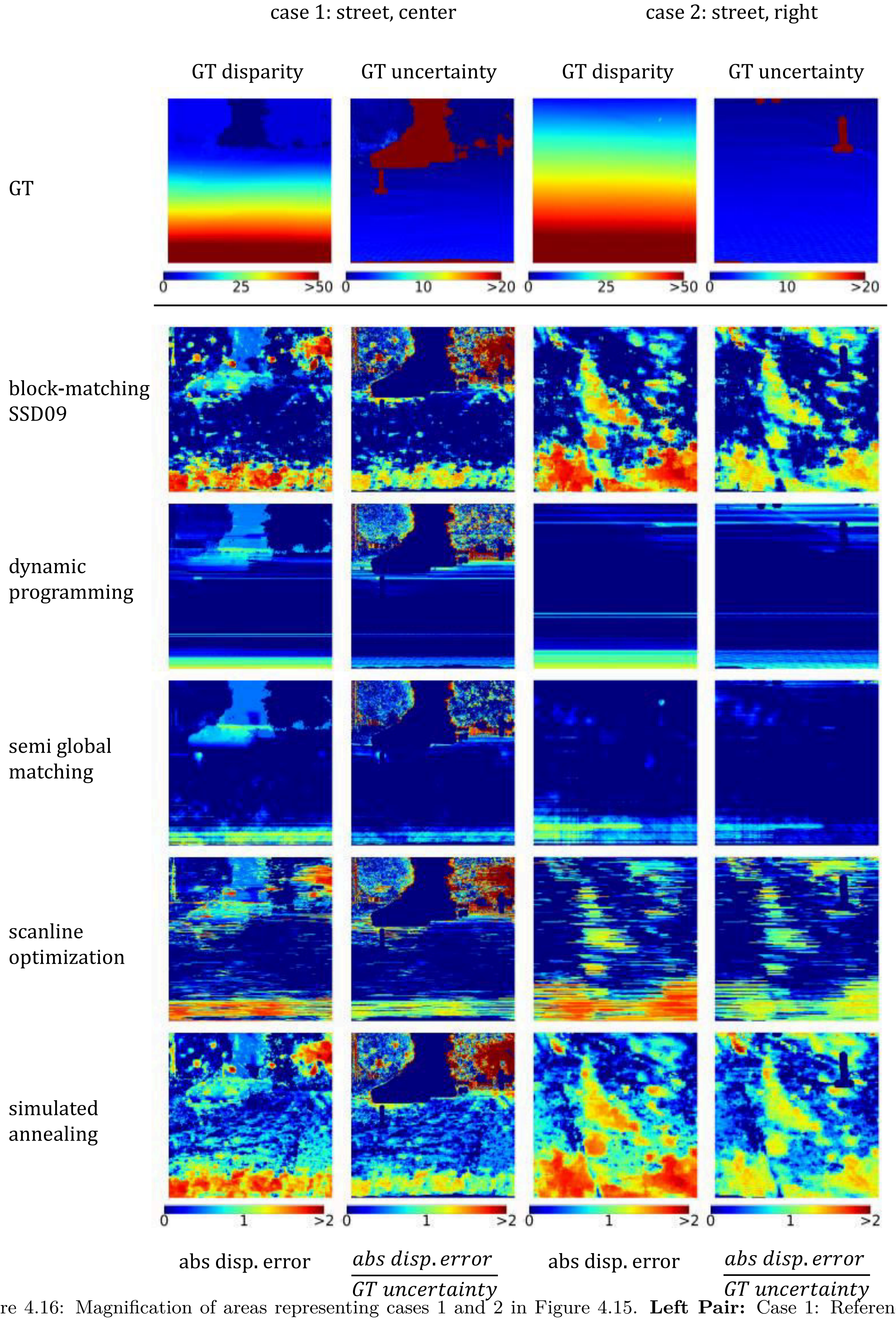


Figure 4.16: Magnification of areas representing cases 1 and 2 in Figure 4.15. **Left Pair:** Case 1: Reference and stereo show strong agreement such that the stereo results are considered very consistent (low values in the weighted image). **Right Pair:** Case 2: Reference and stereo show strong disagreement. Since the uncertainty in this are is low, this is a fail case of the stereo algorithm with high probability.(Courtesy Katrin Honauer for creating the crops)

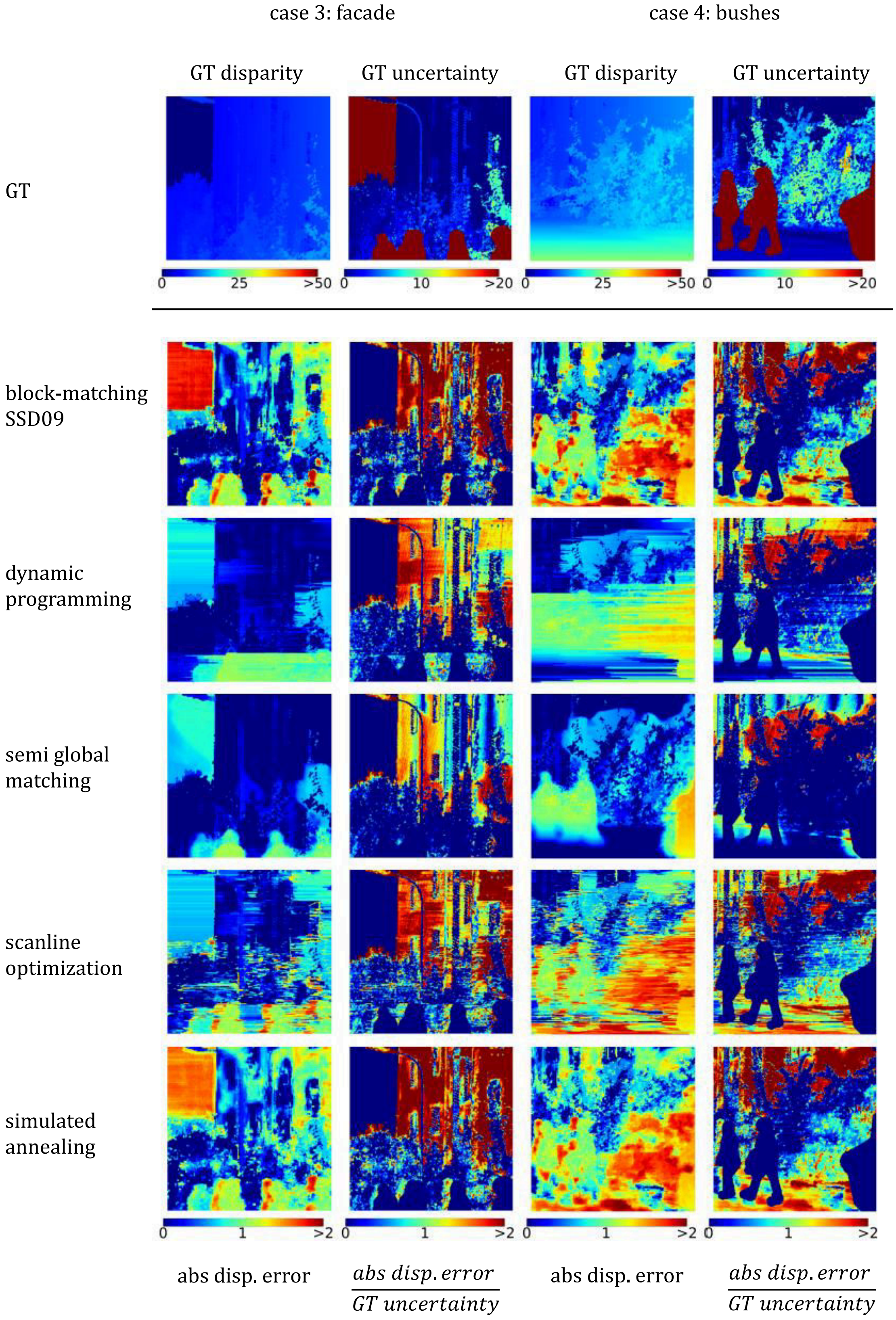


Figure 4.17: Magnification of areas representing cases 3 and 4 in Figure 4.15. **Left Pair:** Case 3: Though the absolute error of the stereo algorithm at the facade is small, the uncertainty allows us to verify that the difference is significant (large weighted values). **Right Pair:** Case 4: While some stereo algorithms have issues with the bushes, the ground truth data is equally uncertain such that we cannot confidently make any statement about algorithm quality without assessing it manually. (Courtesy Katrin Honauer for creating the crops).

the interquartile range at each pixel.

Results with a recorded scene showed that the localization error caused by suboptimal camera estimates significantly deteriorates the quality by introducing multi-modal depth distributions at depth edges, especially at objects close to the camera. Even with arguably the best hardware available today and highly tuned manual alignment tools, the disparity standard deviation exceeds several pixels at nearby objects while simultaneously being less than a pixel for objects with a disparity smaller than 50 piels. Objects with high geometric detail cannot be measured with LIDAR reliably, causing additional artifacts in the ground truth.

Yet, I showed that if the uncertainty of the depth data can be quantified it is still possible to do meaningful performance analysis of stereo data using the reference data and uncertainty distributions.

4.8.2 Outlook

The work presented offers many possible directions for future work. These shall be summed up in the following paragraphs.

Refined Error Model For the proposed method I used the accuracy claimed in the LIDAR manufacturer's data sheet, which should be a very good approximation. In future, more detailed error models as discussed [20] should be incorporated into the analysis.

Refined Experimental Setup In terms of our experimental setup, the accuracy could be improved in smaller scenes by using our approach with a micrometer-accurate structured light scanner delivering object meshes rather than point clouds. Then, the limiting factor becomes camera pose estimation, which is a matter of future studies.

Application to other Acquisition Setups The method presented is not limited to our measurement setup. The techniques can be used, as long as the process of fitting the reference data to the stereo camera frame can be formulated as an energy minimization problem. More specifically, it is possible to apply these techniques to existing stereo reference datasets such as Middlebury and KITTI to supply per-pixel uncertainty estimates. The outcome of such an analysis would either verify the claims on the the uncertainty made the authors quantitatively or give additional insights into where the bounds given do not hold. For KITTI, the applications of the formulas are quite straightforward since their setup was the most similar to ours. Since the LIDAR - stereo setup had a fixed relative pose, the pose uncertainty could be obtained using their calibration data. However, this doesn't necessary result in fixed disparity space uncertainties as they additionally apply a ICP [34] step to aggregate multiple LIDAR frames to one point cloud. Since ICP (once outliers are removed) is a form of least squares fitting, the uncertainty estimate can be easily added to the total minimization functional. For the Middlebury datasets, the process of estimating the pose uncertainty remains largely the same. However, the uncertainty of the reference data has to be obtained in a different manner, as it is based on stereo matching on highly structured objects (by using UV paint). Fortunately, stereo matching can also be interpreted as a least squares problem and the uncertainty of an estimated correspondence can estimated using techniques such as the ones described in [70].

Application to other Vision Problems Similarly, the approach presented is not limited to the generation of reference data with uncertainties for stereo matching. With the depth map aligned with the stereo frame and the relative movement between individual frames in time known, it is a simple matter of reprojecting a depth map into the subsequent frame and measuring the displacement to obtain reference data for optical flow estimation. The parameter fit now not only depends on the current pose estimate, but also on the pose of the subsequent frame. Therefore it is possible to obtain flow uncertainties in a similar fashion by including nut just the pose uncertainty of the next frame Σ_{tj+1} into the error propagation, but also the block diagonal entries that correspond to the covariance between the current pose t_j and the pose of the next frame t_{j+1}. Another, somewhat simpler and faster approach is to propagate the full per-pixel error distribution into the next frame. This second approach was done for 100 frames as a proof of concept and the preliminary results are depicted in Figure 4.18. It is not clear yet how the simplifications made along the way in the simpler approach will affect the reference data quality. However, if they methods are equivalent then the simpler one is obviously preferable. Yet, further

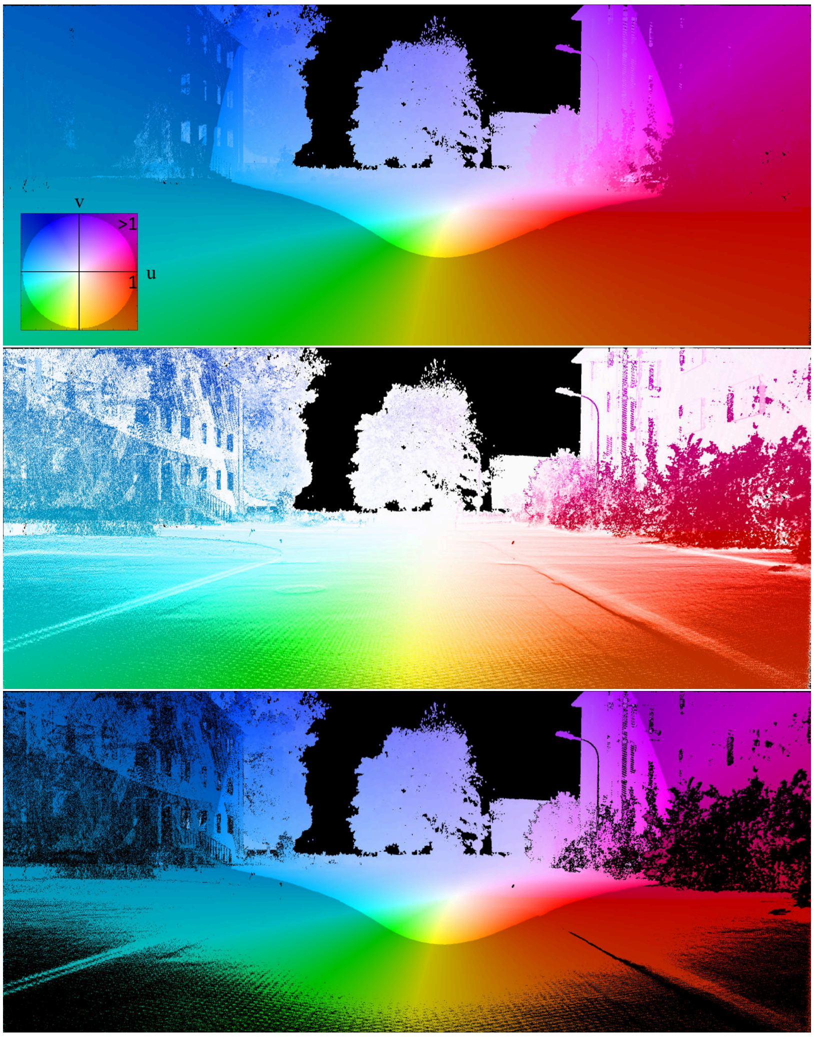


Figure 4.18: Preliminary Optical Flow Reference Data with Uncertainty Estimates. Top: Preliminary optical flow reference data encoded in HSV with a threshold at 1 pixel of flow magnitude (grayed out area). **Middle**: Corresponding interquartile based scalar uncertainty (HSV threshold at 0.1 pixels). **Bottom**: Flow ground truth with masking out regions with uncertainty larger than 0.05 pixels

investigation is required to verify or disprove this supposition.

Application to ToF Stereo Fusion As I argued in the beginning of this chapter, depth map initialization and priors require uncertainties that take misalignment properly into account. To this end, the disparity space uncertainty estimates presented here can be plugged into any fusion technique that makes use of uncertainties.

Bootstrap Alignment A point that can be criticized is the requirement for manual intervention to obtain absolute pose estimates. While the tools present do speed up the annotation process considerably , in practice, it still is not very scalable. Considering that there currently are over 200 sequences that need to be processed and the availability of computing resources, the annotation becomes the bottleneck. Therefore future work on

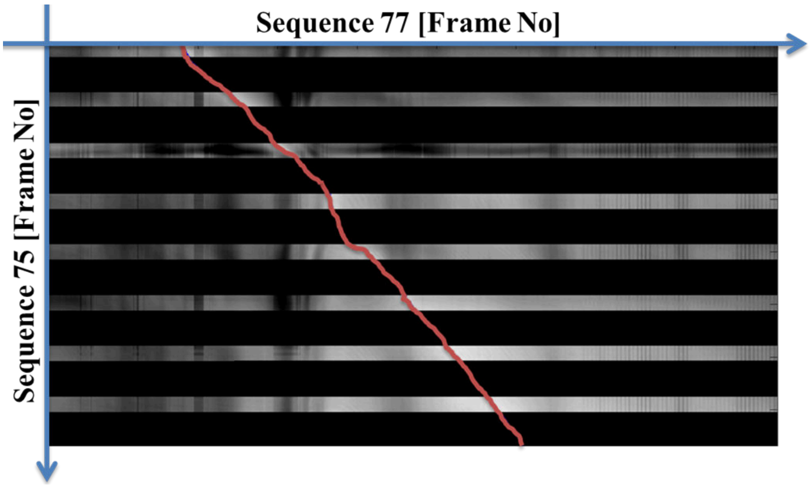


Figure 4.19: Cost Matrix for the bootstrap alignment of sequence 75 (vertical axis) to sequence 77(horizontal axis). The brighter the pixel, the larger the normalized cross correlation between the two frames. The cost was only computed for blockwise equidistant frames in sequence 75. The red line indicates the matching with the largest correlation for each frame. For the black regions, linear inter-/extrapolation was used. The thickness of the line and large resolution of the image mask the fact that the corresponding frames have a certain jitter of up to 10 frames between subsequent frames

further reducing manual intervention is required. One possible solution is to harness the fact that the sequences are similar for similar pathways that were driven. Therefore automatic frame synchronization techniques such as [46] could be used to obtain a rough estimate of camera pose. Automatic Structure from Motion approaches between these frames could then be used to obtain a more precise pose estimate. A proof of concept that this method is potentially useful is illustrated in Figure 4.19. Here, a distance metric is computed between one of the 7000 frames in a unregistered sequence and between all 9000 frames from a registered sequence. More specifically, I computed the normalized cross correlation between the whole images. To speed up computation, this process was done for blocks of frames equidistantly distributed over the whole sequence. Subsequently, the best matching pairs are registered with each other using a greedy strategy. It should be noted that for obtaining the best pairing without jitter etc, the problem can by solved using a dynamic programming papproach similar to dynamic programming stereo matching techniques that work on a single row. Figure 4.20 depicts some example pairings of frames. As rough estimate by visual inspection, the relative distance between bootstrapped and original scenes was at most a couple of meters, thus enabling the usage of structure for motion techniques. Yet again, it should be stressed that this is only a proof of concept. The experiment was undertaken on two sequences that were acquired in close temporal proximity. Therefore lighting, weather and other global scene properties were very similar. Further research is needed to verify these results. Also more elaborate techniques will probably be required to make this approach work in practice considering the different types of scenes.

Key-point free registration A final point I would like to address is the fact that the current pipeline makes use of key-point correspondences for pose estimation. While this is generally considered acceptable, it does remain unsatisfactory from two viewpoints. First, it should be noted that generic automatic key-point detectors are often biased and do not have an isotropic error distribution. The Harris corner detector chosen in the work presented displayed results most unbiased in my experiments, yet a certain remaining bias remains and can be observed as e.g. window corners being consistently picked 1 pixel below the actual position of the corner. The other point is that we are not directly fitting our raw measurements (LIDAR and stereo images) to our model of the world. Instead, first derived quantities are obtained (the key-points) which are subsequently matched. Simply put, it is difficult to directly fit the range data to the stereo frames due to the projective nature of image formation and so far, little work exists on this topic. Yet, preliminary experiments on synthetic data have shown that it may well be possible if an initial guess of pose is close enough to the true solution. To this end, a synthetic stereo dataset with ground truth has been created with similar specifications as the real setup. The relative pose between ground truth and stereo frame was then perturbed by up to 10 cm in translation and 2 degrees in rotation. The algorithm's objective was then to minimize the photo consistency of the projected depth map by rotating and translating the point cloud. By using a combination of random

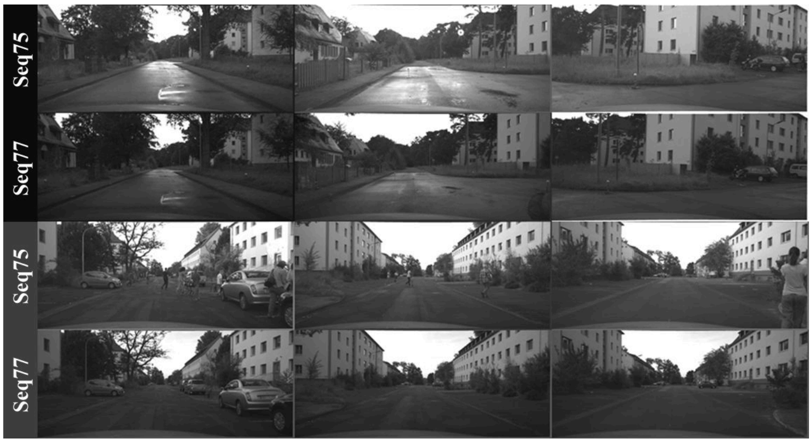


Figure 4.20: Examples of bootstrap alignment between sequence 77 and 75. **Top of image pair**: Sequence 75 that is not registered. **Bottom of image pair**: Registered sequence 77. Results shown from top left to bottom right for frames 62, 1779, 2200, 4108, 5928 and 7004. The other images show similar performance. Visual inspection of these image pairs suggest a relate pose difference bound by 2 - 5 meters.

sampling and gradient descent, it was possible to achieve a similar pose error as stated in Table 4.3. While these results are promising, issues remain that prevent the usage on real data and are thus subject to further research. With an unconstrained minimization energy, the optimization often tries to find a pose that minimizes areas of occlusion, which is frequently not the desired outcome. Furthermore, without constraining the final pose, a runaway optimization could just push the point cloud out the view frustums of both cameras, thus yielding an undefined behavior.

5
Range Extension of ToF Imagers

5.1 Motivation

Time-of-Flight cameras suffer from range ambiguity due to the periodic correlation function, from which the phase is estimated. The disambiguity range can be extended by decreasing the modulation frequency. However, this implies a larger uncertainty in the depth estimate as the uncertainty in phase does not depend on the modulation frequency[1]. The insights gained during the work on Chapter 3 led to the question whether it is possible to resolve data ambiguity using only measurements of the ToF camera without compromising on parameter confidence.

5.2 Contributions

In the following, I revisit the ToF parameter estimation problem and show that the disambiguity range can be greatly extended if different modulation frequencies are used *in the individual subframes*. The resulting least squares problem no longer has an analytical solution. Yet, I show that it is still possible to estimate the desired parameters by combining a grid search and continuous optimization. Furthermore, I produce initial evidence that

[1] for a constant number of cycles of integration

it is possible to speed up the parameter estimation process by initializing the continuous minimization with the output of non-parametric regression.

As a result, I present a simple method towards extending the effective range of a phase based Time-of-Flight (ToF) camera by means of different modulation frequencies in the individual sub-frame measurements. Unlike related work, the proposed method does not rely on strong prior assumptions or additional measurements. At the same time it does not have to bargain on parameter confidence. Finally, to validate my claims, I present results on two real scenes.

The remainder of this chapter is organized as follows: I commence by presenting the related work on range extension in Section 5.3, before giving a phenomenological treatment of the ToF measurement model with multiple modulation frequencies in Section 5.4. Subsequently, I present experiments on solving this model and initial results on parameter initialization in Section 5.5. Finally, Section 5.6 concludes this chapter with a summary and outlook on future work.

5.3 Related Work

The simplest way to extend the range of ToF cameras would be to choose a lower modulation frequency as this naturally corresponds to a larger disambiguity range. Yet, remember that the reconstruction is based on estimating a phase and the uncertainty in the phase estimate remains unchanged. This leads to an increase in the uncertainty of the depth estimate. Another issue is that for a similar signal to noise ratio, one would require longer integration times leading to stronger motion artifacts.

Therefore, methods exist that try to extend the range at higher modulation frequencies. Three different classes of such methods can be identified. Following [74], one can distinguish between methods that use a single depth map obtained from the ToF camera and methods that use information from multiple depth maps. Both these classes are oblivious to the ToF measurement principle, i.e. they do not require access to the raw data. This is in contrast to the third class of methods that operate directly on the raw data. The proposed method also belongs to this last class.

Single-frame based techniques make use of a combination of the following cues:

- The modulation source can often be approximated by a point light source. This leads to a quadratic falloff of signal amplitude with distance. By making the assumption that the scene reflectance is constant over the whole image, it is possible to distinguish between different cycles of the phase [147, 36].
- The other Ansatz is to introduce a regularizer that penalizes depth discontinuities of more than the disambiguity range [63, 57].

Both techniques make strong assumptions on the composition of the scene. The first assumption is violated in presence of objects with different infrared albedos. Similarly, large discontinuities cannot be handled by the latter approach. Additionally, it is not possible to reconstruct absolute depth if all objects present are offset by more than one ambiguity range.

Methods based on multiple depth maps acquire two depth frames $\mathbf{R}^1 = \{r_i^1\}$ and $\mathbf{R}^2 = \{r_i^2\}$ at different modulation frequencies and then seek two numbers $n_{1,i}$ and $n_{2,i}$ for each pixel i such that

$$r_i^1 + n_{1,i} \cdot r_{amb}^1 = r_i^2 + n_{2,i} \cdot r_{amb}^2, \tag{5.1}$$

where r_{amb}^1 and r_{amb}^2 correspond to the ambiguity ranges of the two modulation frequencies. The two depth images are acquired either sequentially with the same camera [51] or simultaneously using a multiple camera setup [35]. In sequential approaches, the integration time is doubled, thus reducing the effective frame rate and introducing additional motion artifacts. The simultaneous approach, on the other hand, requires additional hardware and has to consider calibration issues between cameras.

The final class of methods, including the proposed one, are based on the raw data parameter estimation problem. [143] present an approach to range extension, where a sum of two sine waves is used as the modulation signal. This is emulated by sequentially emitting two different sine modulated signals during a *single* integration period. Then, using Fourier analysis the depth is reconstructed from at least five sub-frame measurements. In contrast, for the method proposed four measurements suffice. Moreover, it can be used with little modification of current hardware.

5.4 Measurements with Multiple Frequencies

5.4.1 Parameter Estimation Revisited

Let us recall the raw image formation process (Eq. (2.33)) of each subframe. The ith subframe measurement in pixel j is modeled as

$$I_j^{T,i}(g_j, a_j, \phi_j) = g_j + a_j \cos\left(i\frac{\pi}{2} + \phi_j\right). \tag{5.2}$$

In the following, the index T and j will be omitted for clarity. Given the measurements I^i, the reconstruction formula for ϕ is given by

$$\phi = \operatorname{atan}\left(\frac{I^3 - I^1}{I^2 - I^0}\right). \tag{5.3}$$

Remember that this is the closed form solution that - together with the other Equations (2.37) - minimize the least squares energy

$$E(g, a, \phi) = \sum_{i=0}^{3}\left(I^i - \left(g + a\cos\left(i\frac{\pi}{2} + \phi\right)\right)\right)^2. \tag{5.4}$$

Due to the cosine in Eq. (5.4), the residual energy is cyclic in ϕ such that for any $k \in \mathbb{Z}$

$$E(g, a, \phi) = E(g, a, \phi + 2\pi k). \tag{5.5}$$

Since the depth r is related to ϕ by the speed of light c and modulation frequency f_m

$$r = \frac{c}{4\pi f_m}\phi, \tag{5.6}$$

the residual energy is also cyclic in r with periodicity of

$$r_{amb} = c/(2f). \tag{5.7}$$

This is illustrated in the top left image in Figure 5.1, which plots the residual energy for true depth vs. estimated depth for fixed intensity and amplitude.

By reparametrizing Eq. (5.4) using Eq. (5.6) and replacing the single modulation frequency f_m with a subframe dependent one $f_i, i = 0...3$, we obtain:

$$E(g, a, r) = \sum_{i=0}^{3}\left(I^i - \left(g + a\cos\left(i\frac{\pi}{2} + f_i\frac{4\pi}{c}r.\right)\right)\right)^2. \tag{5.8}$$

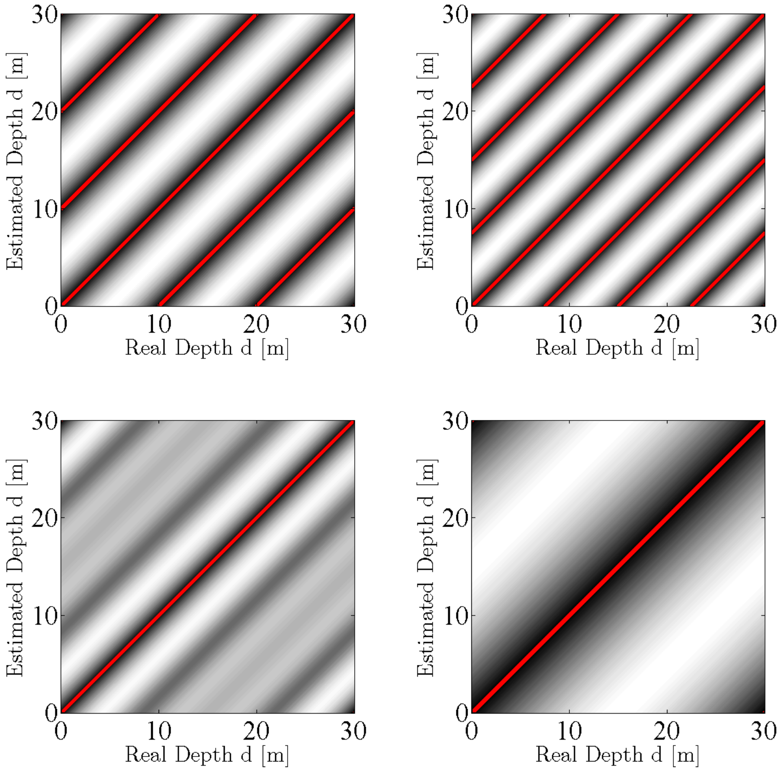


Figure 5.1: Energy Surfaces For Different Choices of f_i. Intensity and Amplitude of the true signal were kept constant. The true depth was varied along the x axis while the cost for estimating a certain depth color coded along the y axis. Light color means high cost, dark color means low cost. Red lines indicates regions with the lowest cost. **Top**: Single frequency across all subframes a) 15 Mhz and b) 20 Mhz leading to a disambiguity range of 10 m and 7.5 m respectively. **Bottom**: c) Combination of 15 Mhz and 20 Mhz yields a ambiguity range of 30 m. d) This corresponds to a single frequency measurement at gcd(15,10) = 5 Mhz. Note the wider lobes in d) as compared to c). Also note the suppressed side minima in c).

Note that this energy no longer has a simple closed form solution in general. In the following, I will instead give a phenomenological presentation of the structure of the problem.

5.4.2 Qualitative Assessment

Energy Surfaces In the following, I consider the residual energy surfaces for true depth against estimated depth for fixed intensity and amplitude for various choices of modulation frequencies. I first compare the energy surfaces resulting from choosing two different frequencies $f_0 = f_1 = \nu_0$ and $f_2 = f_3 = \nu_1$ with the energy surfaces using $f_i = \nu_0 \, \forall i$ or $f_i = \nu_1 \, \forall i$ respectively (cf. Figure 5.1 a) -c)). With two different frequencies, the effective range in which the global minimum of the energy is unambiguous is extended to the greatest common denominator of ν_0 and ν_1, similar to the effect seen in the related work using two depth maps obtained at different frequencies. Comparing the two-frequency-measurement with a standard measurement using the greatest common denominator (cf. Fig 5.1 c) and d)) confirms that both frequency combinations yield the same disambiguity range. Note that the width of the lobe containing the global minimum is much narrower when using two frequencies.

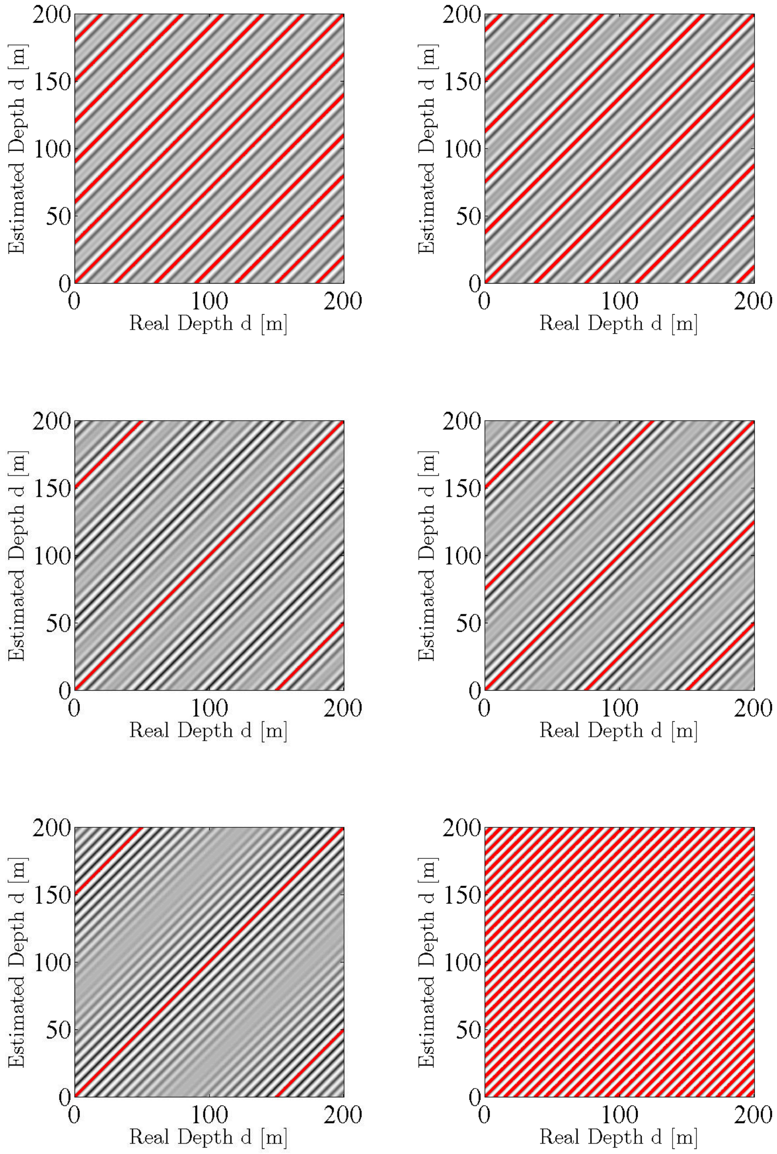


Figure 5.2: Combining Two Frequencies: **Top left to bottom right**: 20 Mhz combined with a) 15 Mhz, b) 16 Mhz, c) 17 Mhz, d) 18 Mhz, e) 19 Mhz and f) 20 Mhz. With the frequencies closer to each other, the disambiguity range is extended while at the same time the strength of the side minima is enhanced (Making it more challenging to find the global minimum).

This corresponds to a smaller uncertainty of the parameter estimate and therefore, assuming the same noise characteristics, also corresponds to a more robust depth estimate.

The observations up till now would suggest choosing f_0 and f_1 as close as possible and as large as possible to maximize the disambiguity range whilst minimizing parameter uncertainty. This is because I have largely ignored the existence of other local minima in the range of disambiguity. Choosing the largest frequency possible will yield a high periodicity of the local minima and choosing the frequencies close to each other will cause the minima caused by ν_0 and ν_1 to amplify each other. If noisy

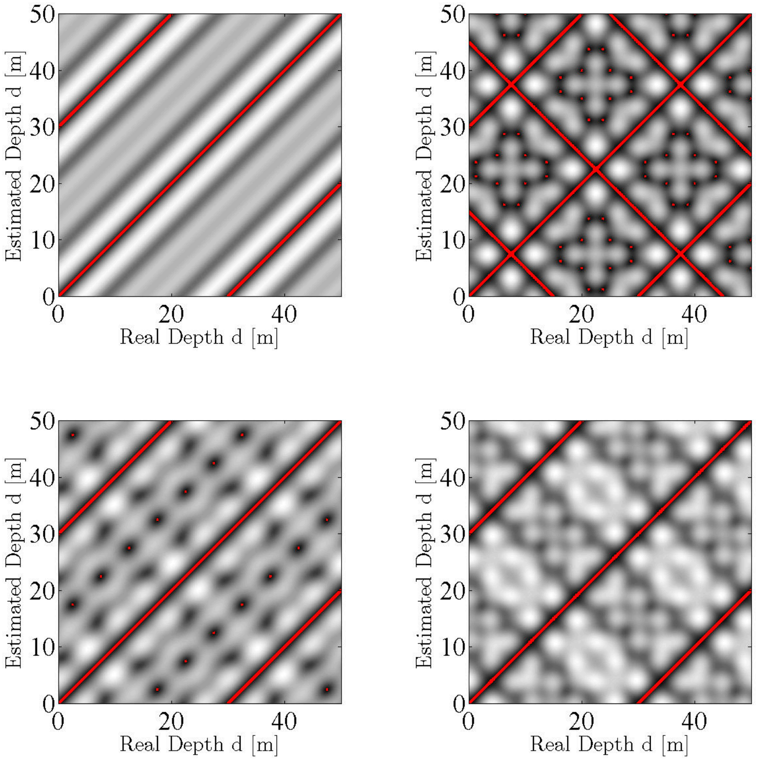


Figure 5.3: Energy Surfaces for Different Combinations of Two and More Frequencies. **From top left to bottom right**: Energy surfaces for a) using two different frequencies sequentially, b) using two different frequencies alternately, c) using three frequencies and d) using four frequencies.

measurements are considered and the theoretically global optimum is not distinct enough from the side minima, this will lead to completely erroneous parameter estimates. This is illustrated in Figure 5.2, where $f_0 = f_1 = 20\,\text{Mhz}$ and the other two frequencies varied between 15 and 20 Mhz. Note how the range is extended until the greatest common denominator is 1 Mhz (c) and e)), while at the same time the side minima become stronger. Therefore, a tradeoff has to be made between large theoretical ambiguity and confidence in the fitted parameters and the suppression of side lobes. The choice can then be additionally constrained by possible hardware limitations.

Alternating the two frequency measurements instead of measuring them sequentially[1] roughly doubles the amount of additional local minima while at the same time giving the local minima lower energy. Therefore, empirically, the sequential case seems to be better.

The insights gained above extend to three and four measurements (cf. Figure 5.3), with the effective disambiguity range being the smallest common denominator of all available frequencies. With each additional frequency, the width of the main lobe containing the minimum is narrowed further. At the same time,

[1] equivalent to $f_0 = f_3 = \nu_0$ and $f_1 = f_2 = \nu_1$

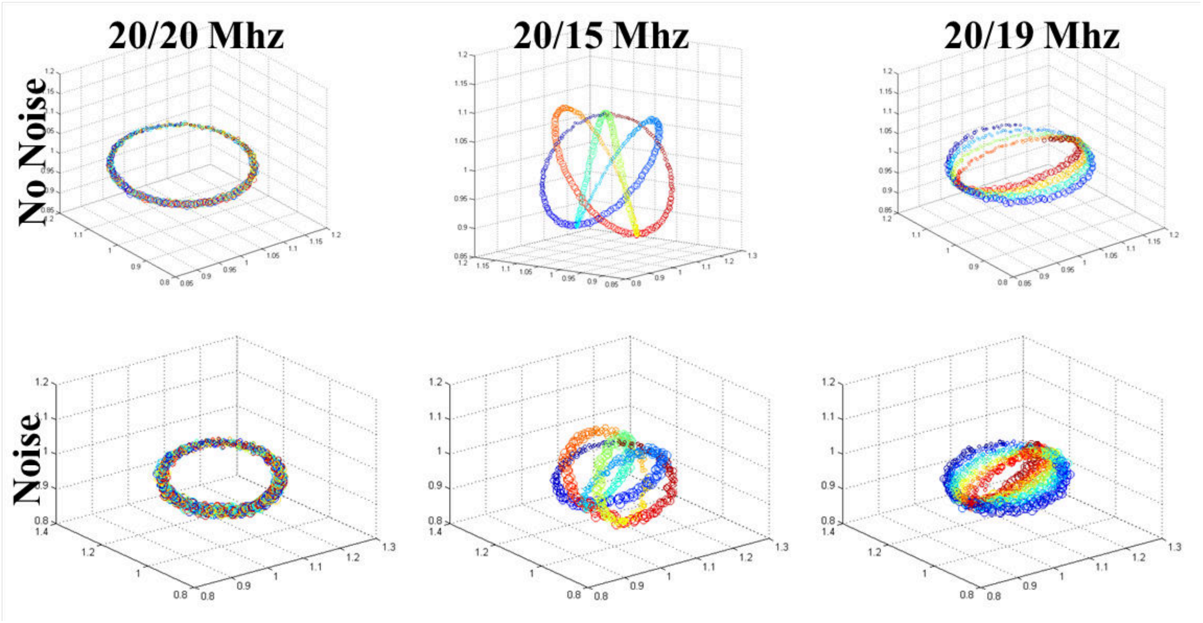
additional minima are introduced.

Trajectories The analysis of the trajectory of the raw data measurements as a function of true depth gives further insight into why using multiple frequencies is beneficial. In Figure 5.4 we visualize the mapping

$$(I^0, I^1, I^2, I^3) \rightarrow d \tag{5.9}$$

as a scatter plot. $I^0 - I^3$ are plotted against the x, y and z axis. I^4 corresponds to the marker size and the depth corresponds to the color. For fixed amplitude and offset, the depth was varied from 0 - 30 meters. A small amount of jitter was added to the markers to identify ambiguous regions. In the standard case, we obtain a closed periodic trajectory where $R_0 - R_4$ map to multiple depths. By choosing 15 Mhz as the second frequency, the curve corresponds to a multidimensional Lisajou figure without intersections (in four dimensions), such that each depth corresponds to a distinct combination of raw measurements. The closer the two frequencies are, the tighter the trajectory gets, such that a slight perturbation of raw measurements leads to a point in 4D space that actually belongs to another depth. Understanding that there is a bijection between raw data and true depth and that the trajectory is smooth, motivates the usage of non-parametric regression techniques to obtain an initial guess for the global optimum.

Figure 5.4: Visualization of the Mapping from Raw Data to Real Depth. **Top:** Little noise and **Bottom:** More noise in the input. **Left:** Single frequency (20Mhz). **Middle:** 20 Mhz and 15 Mhz. **Right:** 20 Mhz and 19 Mhz. The raw data are mapped to the 3 spatial axis and the marker size. The depth was varied from 0 to 30 meters. The corresponding true depth is color coded. For a single frequency case, the trajectory corresponds to a circle with many ambiguity cycles (amounting in the noisy visualization). Using two frequencies, the trajectory corresponds to a Lisajou figure such that the true depth remains unambiguous in a larger range. Choosing two frequencies too close to each other (right side) causes the trajectory to pass regions that are close by such that even little noise can cause the wrong depth being estimated.



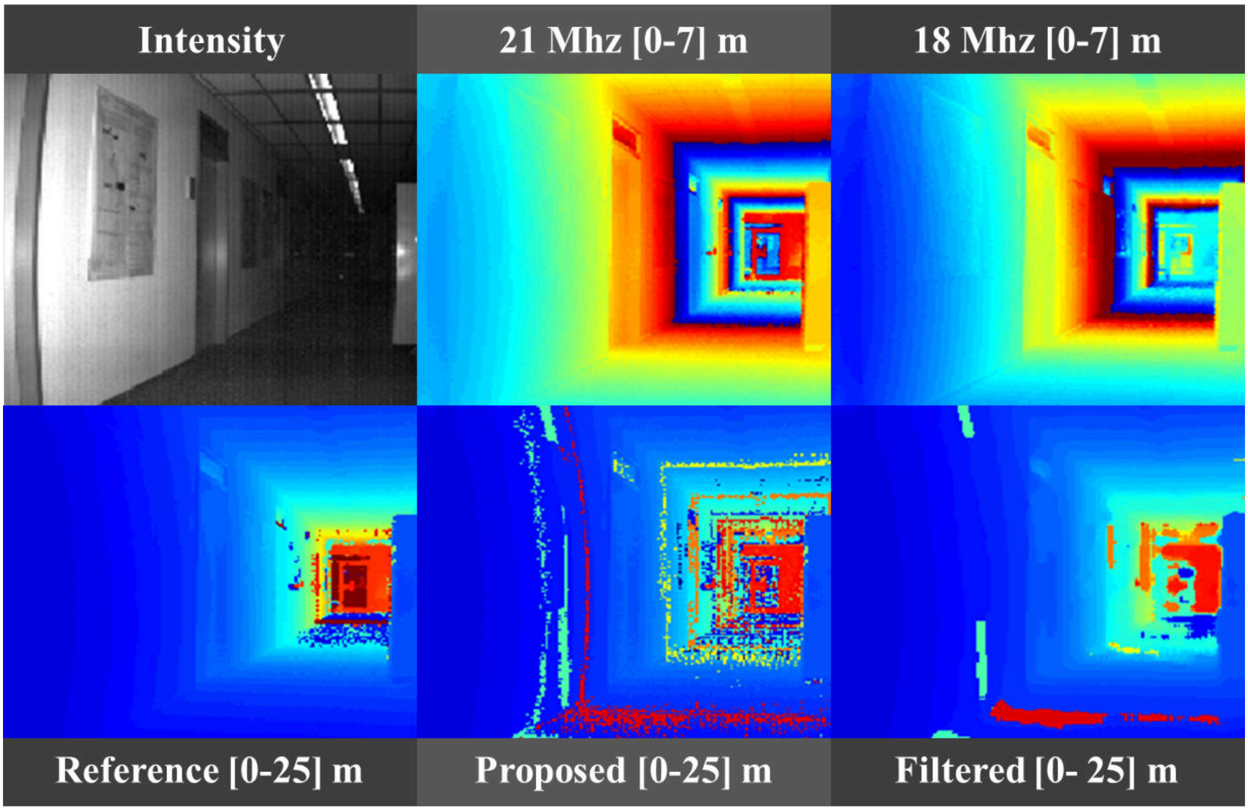


Figure 5.5: Result of Range Extension on Scene 1 (Hallway). **From top left to bottom right:** a) Intensity, b) Depth image at 21 Mhz, c) Depth image at 18 Mhz, d) Reference data set obtained by applying Eq. 5.1 to the first two depth maps, e) Result of the proposed method using only 4 sub-frames at 21 and 18 Mhz and f) result after post-processing with a 5 × 5 median filter. We see that it is possible to extend the range using only 4 measurements at two high modulation frequencies. Another thing we can observe is the depth error at periodic intervals. This might be due to the grid used for initialization of continuous optimization being to coarse.

5.5 Experiments and Results

Due to the additional local minima, a line search or trust regions method was combined with a grid search to solve Equation (5.8). In the following, I limited myself to choose between four frequencies, i.e. 18, 19, 20 and 21 Mhz, as they correspond to the possible frequencies that the available camera[2] can be set to. This yields a total of 256 possible frequency combinations.

The best combination of frequencies was determined by eyeballing (for a disambiguity range of up to 30 m). Here, $f_0 = f_1 = 21$ Mhz and $f_2 = f_3 = 18$ Mhz. To make the system work on the real camera system, Equation (5.8) had to be extended to compensate for differences in camera gain at different modulation frequencies. The gain factors were obtained by linear approximation of the internal camera calibration. The results are depicted in Figures 5.5 and 5.6. The scene depicted in Figure 5.5 is a hallway containing various different albedos and a maximum depth of 30 meters. In Figure 5.6, we observe a large courtyard scene where the minimum depth is already beyond a single cycle of the single frequency measurement. For com-

[2] PMD Camcube 3

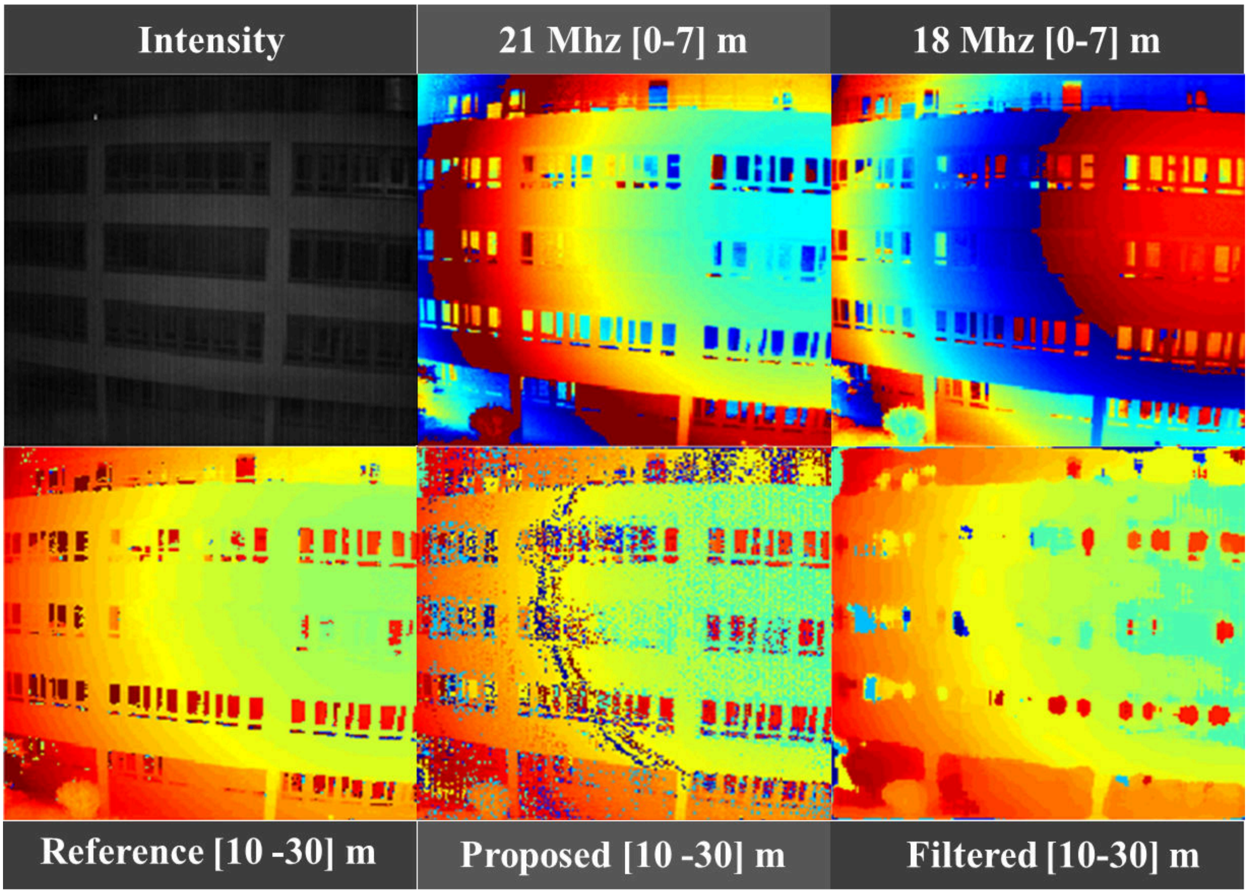


Figure 5.6: Result of Range Extension on Scene 2 (Courtyard). **From top left to bottom right:** a) Intensity, b) Depth image at 21 Mhz, c) Depth image at 18 Mhz, d) Reference data set obtained by applying Eq. 5.1 to the first two depth maps, e) Result of the proposed method using only 4 sub frames at 21 and 18 Mhz and f) result after post-processing with a 5 × 5 median filter. Here, we see another benefit of the proposed method over traditional single frame unwrapping methods: If all depth values are shifted by one or more cycles, these methods cannot recover the absolute depth.

parison, I display the range maps obtained by the individual frequencies, a ‘reference image’ obtained by applying Eq. 5.1 to two full frame measurements as well as the depth map obtained by the proposed method and a 5 × 5 Median filtered version thereof. We can see in both scenes, that the proposed method is able to extend the range of the ToF system. Yet, there are erroneous measurements visible at specific distances due to erroneous local minima. I believe that these may be distances at which the trajectory of the raw data may be too close to another local minimum (cf. Figure 5.4). Note also that the scene in Figure 5.6 is such that common single-frame based methods will not be able to reconstruct the absolute depth.

Regression Doing a grid search for obtaining the global minimum is not feasible for systems designed to deliver data in realtime. The evidence in Figure 5.4 showed that, while no analytical solution can be obtained, it might still be possible to learn the mapping between raw channels and depth. This mapping

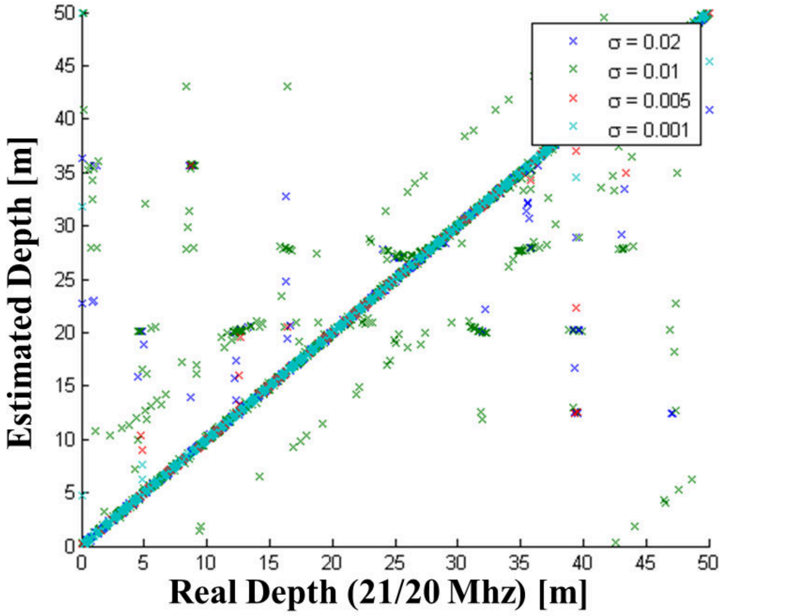


Figure 5.7: Result of Non-Parametric Regression for Initialization of the Continuous Optimization. The mapping between the raw data channels and the true depth was learned from synthetic reference data sampled on a grid. The method used was an ensemble of regression trees. Here I present the predicted depth values given input data perturbed with different levels of noise. For comparison: For the real camera employed, the noise level (normalized by maximum raw data value) is around 0.002-0.006 (for scenes with normal indoor lighting).

was learned for $f_0 = f_1 = 21\,\mathrm{Mhz}$ and $f_2 = f_3 = 20\,\mathrm{Mhz}$ [3] using regression tree ensembles [26]. The input data was generated synthetically on a grid with 10 cm spacing in the distance between 0 and 50 meters. The amplitude chosen at a spacing of 0.02 between 0 and 0.2 and the intensity was chosen at a spacing of 0.1 between 0 and 1. After inference, the locations of the parameters were estimated using the synthetic raw data with different noise levels as input. The results are depicted in Figure 5.7. As we can see, it is possible to learn an initial mapping between raw data and the true depth. The majority of points lay in the correct local minima for all levels of noise. For the two larger noise levels, the estimates sometimes fall into the next local minimum. Yet, even this case can considerably reduce the search space of initial values for continuous optimization.

5.6 Summary and Outlook

5.6.1 Summary

I this chapter, I presented a simple method capable of extending the disambiguity range of Time-of-Flight cameras by controlling the modulation frequency between sub-frames. I commenced by providing a phenomenological overview of how the choice of dif-

[3] This amounts to a disambiguity range of 50 m.

ferent modulation frequencies affects the energy surface of the least squares problem and the mapping between raw data channels and depth. Next I showed on real datasets that the resulting least squares problem can be solved using a combination of grid search and continuous minimization. Finally, I provided evidence that the optimization can be sped up by means of non parametric regression to obtain an initial guess.

5.6.2 Outlook

In the future, two lines of work are of particular interest:

Sensor Model As I mentioned in the previous section, everything that affects the mapping from raw data to metric depth has to be known for this method to work. For real cameras, this also means that the sensor response and internal depth calibration have to be known analytically. The results I presented used a simple linear approximation to the unknown internal calibration of the utilized camera. Instead, future work should focus on more refined analytical models [163] to approximate this mapping.

Experimental Design Currently available cameras do not provide the additional hardware control required to arbitrarily choose modulations frequencies. Yet, this is not an inherent limitation of the hardware[4]. Therefore, assuming that there is a large range of possible frequencies, the question comes up how to choose the best ones. This corresponds to a well studied question in the field of experimental design. Applying the techniques known in that field could therefore yield interesting answers of what design parameters a Time-of-Flight camera must have to achieve the smallest measurement uncertainty over a wide range of parameters.

[4] Based on personal correspondence with manufacturers.

6 Reflections on Stereo

6.1 Motivation

Specular surfaces are a source of error for many depth imaging techniques, be it ToF imaging and passive stereo as we have seen it in Chapter 3 or other methods such as structured light. It is therefore essential to a) understand how reflective surfaces cause errors and to b) investigate, whether such errors can be accounted for in a principled manner. In the following, I present work on the analysis and modeling of reflections for passive stereo reconstructions.

The traditional approach to stereo matching frequently models image formation as a world consisting of Lambertian surfaces observed through a perfect pin hole camera. While these assumptions, together with the right regularization, do suffice in many settings, there remain real-world situations where this is not the case. Recent benchmarks and challenges [124, 60, 187] have shown that there are often situations where the imaging model is violated, whether geometrically or radiometrically (e.g. different gains, non-Lambertian surfaces, lens flare). Reflective surfaces violate the Lambertian world assumption and cause the observed color of a surface point to depend on the viewpoint. In turn, this leads to false minima in stereo matching data terms that depend on some form of brightness constancy (cf. Eq. (2.25) and Figure 6.1). The traditional approach to handle specular

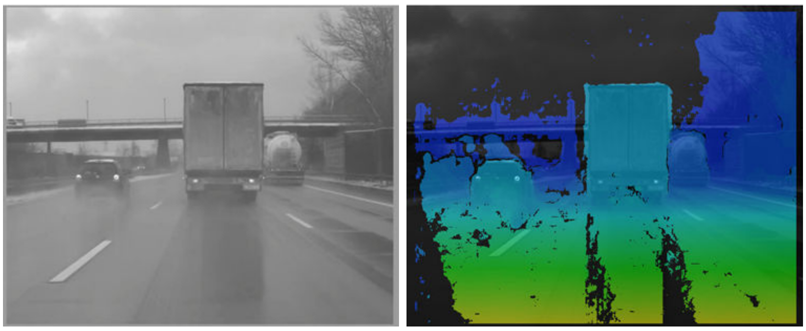


Figure 6.1: Illustration of Errors Caused by Reflective Surfaces. **Left:** Stereo image. **Right:** disparity map resulting from the RankSGM stereo method [77]. The color observed on the road at the reflection of the silhouette depends on the position of the camera. The algorithm therefore assigns erroneous disparity values at this location.

surfaces is either by robust data terms (e.g. correlation or rank order statistics) or by using strong regularization techniques. The work presented here has a different goal and was guided by the following questions: Is it possible to derive a data term that explicitly takes reflective materials into account? And if so, is this model of any use? Can we estimate scene parameters using this model?

6.2 Contributions

The findings on stereo with reflections are summarized in Figure 6.2. By additionally modeling up to two bidirectional reflectance distribution function (BRDF) parameters (cf. Section 2.1.2), it is not only possible to remove errors due to reflective surfaces, but it is also possible to obtain material information from the two images that can potentially be used for segmentation purposes or view synthesis. Finally, while not explicitly estimated, the separation of diffuse components and reflection components falls out of the box. Note that the work presented does not include any form of global regularization or post-processing on top of the presented results. This derived from the goal is to give insights upon the utility of the proposed models. The images displayed are a sole result of the proposed models and per-pixel inference techniques.

In Section 6.4.2, I revisit the roots of stereo matching as a least squares problem and from this formulation derive simplified models that take reflections into account. Also, I show that traditional diffuse world stereo is in fact just another special case. The models are parameterized by per pixel depth, normals and up to two surface material parameters which encode strength of the reflection component and roughness of the reflecting surface. All in all, there are up to 5 parameters per pixel. The resulting optimization problem is high dimensional and requires that the

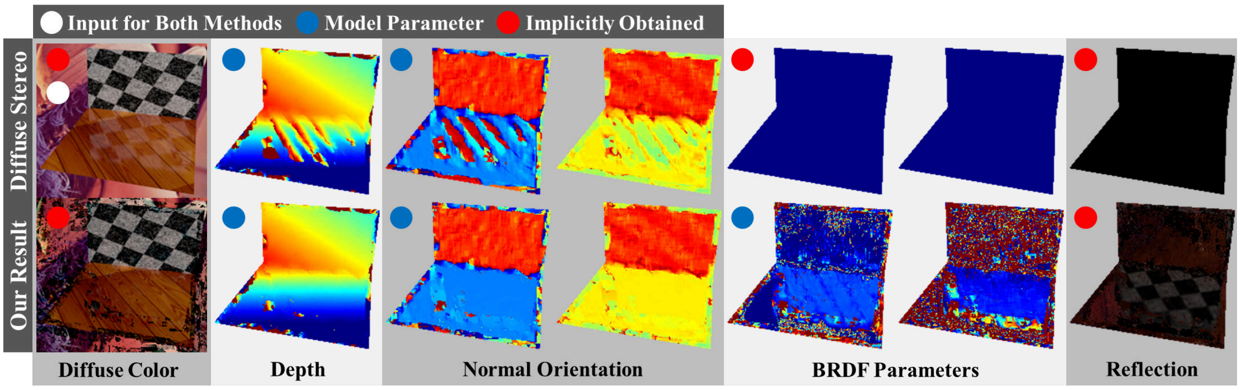


Figure 6.2: Stereo with Reflections. By explicitly parametrizing two BRDF parameters as well as geometry (a total of 5 parameters per pixel), and exploiting the ability of PatchMatch to efficiently optimize high-dimensional energies, it is possible to obtain the BRDF *and* a better geometry.

surface belonging to each pixel 'knows' which surfaces it reflects from, thus yielding large non-local interactions. I demonstrate that the inference still is tractable using Stereo PatchMatch [19] with extensions that enable efficient reflection computation and more accurate normal estimation (cf. Section 6.5.2). While the computation of accurate surface orientation is usually not the main goal in stereo matching, it turns out that accurate normal estimation is the key to handling reflective surfaces. These insights and the properties of the resulting algorithms are further discussed in Section 6.6. Before describing the method I will first review the related work and position the contributions made with respect to them.

6.3 Related Work

Non-Lambertian Vision Early approaches in handling specular surfaces involved the detection of specular highlights [9, 27, 61, 105, 140] and subsequent exclusion of the detected areas from stereo matching. Another approach with a similar goal is the usage of a cost function that is invariant to specular highlights, for example the image gradient [19, 139] or rank-based costs [78, 79]. Both of these cost types achieve robustness towards global or low frequency radiometric differences in the input images, but still have issues with strong specular highlights or high frequency reflections. For handling stronger highlights, Jin et. al. [89, 88] make use of a rank-based cost, though in a multi-view setting. All these methods have in common that they do not change the diffuse world model. Instead, the reconstruction is limited to the diffuse parts, either directly by detection, or indirectly by incorporation in the cost function.

On the other hand, the apparent movement of specular highlights provides information on the normals and surface curvature of object surfaces [17, 195]. These movements have been harnessed to reconstruct mirror surfaces [153, 2, 107]. These methods are in some sense complementary to the methods described earlier as they model perfect mirror surfaces and require controlled lighting conditions and assume orthographic projection.

An approach to treat both reflective and diffuse surfaces — and transparent objects as well — is *layer separation*, where the world is treated as a set of semitransparent depth layers mixing the color with each other. Levin and Weiss [113] use user scribbles of edges belonging to different layers and regularization based on natural image statistics to obtain such a layer separation.

With multiple calibrated views, epipolar plane analysis [21] can be used to separate different depth layers [39, 183]. Finally, for stereo images, Tsin et. al. [180] find multiple layers by 'nested plane sweeping', essentially extending the disparity search space to pairs of depth hypotheses per pixel. These previous models restrict the source of reflected light to so-called 'horizontal' disparities, In reality however, but reflected light can come from anywhere in the scene.

The model presented here is quite different in that the physical properties of the observed surfaces are modeled. These physical properties implicitly define a second observable layer. By doing so, it is possible to use reflection information in the image wherever it is available and thus obtain something close to a material segmentation of the image for free.

An alternative might seem to be an example-based approach to material modeling [178]. Here, the correct matching color values of reflected surfaces under different viewing and lighting angles for a single light source were learned for a few materials. This look-up table was then used as a stereo-matching cost in scenes containing the learned materials. Although this method is similar in that it tries to get material specific information, it differs in that it does not learn the material BRDF itself, but the appearance under a single light source of fixed strength. The number of examples required to learn the case of general reflections would therefore seem prohibitively large using this approach.

Inverse Rendering Finally, the work presented here is closely related to various inverse rendering problems. Here, the world and lighting parameters consistent with an observed image need to be found such that they satisfy the rendering equation [92]. Common problems [142] require that a certain subset of variables has to be known. Inverse lighting requires known geometry and reflectance to estimate light sources [118, 90], and inverse reflectometry is concerned with BRDF measurements with known geometry and lighting [119]. For estimating geometry, Liu and Cooper [115] showed that MRFs can be used with very high order interactions to solve an inverse ray tracing problem albeit still with diffuse reflectance. In [189], the authors showed that the simultaneous estimation of geometry and specular reflectance is possible if the light sources are known. The method described here is similar with respect to the fact that my model formulation has similarities with the Radiosity equation [65] used in many inverse rendering algorithms, though unlike the other methods the goal was to estimate all model parameters jointly.

Inference Techniques The inference problem that is required to be solved here has a high-dimensional state vector at every pixel and an energy with long-range interactions between pairs of pixels. Moreover, the variables participating in the interactions are themselves a function of the unknown parameters. Until recently, this would have appeared tractable only for very simple greedy algorithms. However, recent work on the PatchMatch algorithm [81] has shown that it is an effective optimizer even for very high-dimensional state vectors.

6.4 Reflections on Stereo

Before diving into the model derivation in Section 6.4.2, I would first like to give an intuition on the appearance of specular reflections and how they affect the stereo reconstruction. In the following, I utilize the geometric optics approximation of light propagation, as it suffices to explain the effects under consideration.

6.4.1 Understanding Reflections

Types of Specular Reflections As specular reflections are best understood by example scenes, I want to point to Figure 6.3,

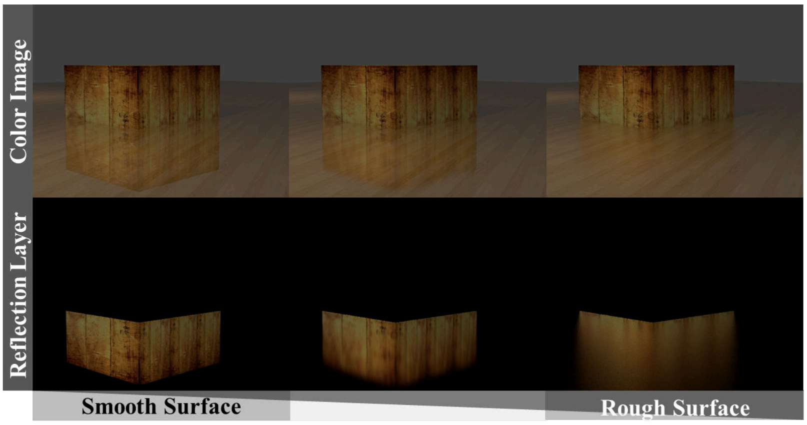


Figure 6.3: Example Scenes with a Specular Floor Surface. **Top**: Scene. **Bottom**: Additive viewpoint dependent reflection component. The relative strength of the reflection is kept constant while the specular roughness is increased from left to right. Note, the low-pass effect that increasing material roughness has on the reflection component. For stereo matching, the mirror reflection case on the left is more difficult to handle due to the additionally introduced color edges by the reflections.

which depicts three scenes that were created using a renderer that models the global light transport in the scene. The material properties of the floor were varied while the geometry of the scenes remained identical. The left column depicts a surface that reflects part of the received light like a perfect mirror. This is often the case for coated objects or objects polished to a high degree. The surface normals all point in exactly the same direction such that we can observe a sharp mirror image of the reflected object. First thing to notice is that reflections are an additive property, that is, if the camera has a linear intensity response, the observed image is an addition of a specular component and a diffuse component. Stereo matching works well on the diffuse component while the specular component suggests an erroneous depth that corresponds to the virtual distance of the reflection. With an increased amount of imperfections on a microscopic scale that lead to a distribution of possible surface normal directions on the visible scale, the reflected image becomes more blurred (middle column). This is because the surface increasingly reflects not only from the mirror direction, but also from its surroundings. Most everyday surfaces display this kind of behavior (e.g. in practice it is difficult to perfectly polish a surface). Note how the reflection component is blurred even more with increased micro-facet roughness (right column) such that it is hard to discern the reflection at all if the reflected surface is further away from the observed point. The roughness parameter acts as a distance dependent low-pass filter of the reflected image. This is an important insight, as a model trying to handle reflections should therefore also account for this effect. Also note that stereo methods making a diffuse-world assump-

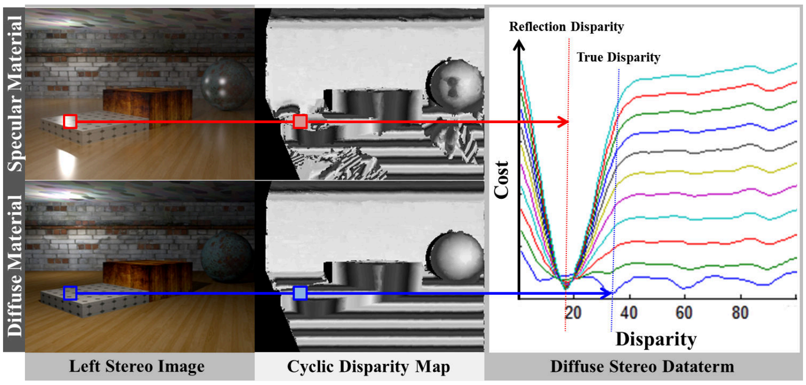


Figure 6.4: Effect of Reflection Strength on Diffuse Stereo Reconstruction. The cost function for assigning a certain disparity to the marked pixel is plotted for increasing strength of the reflection component. Left image and corresponding disparity maps in cyclic coding for a diffuse and a highly specular scene are depicted in the two left rows. Note how the cost function minimum jumps from the right disparity value and the disparity of the reflected image with increased reflection strength.

tion will have little trouble in this final case since most of the signal variation which is used for matching purposes stems from the surface itself. This is why diffuse-world stereo often still works even though most materials do violate the Lambertian world assumption to some degree.

Effect of Reflections on Diffuse Stereo Diffuse stereo methods do not always fail in presence of reflections as illustrated in Figure 6.4. Whether the right distance is estimated depends on the relative strength of the reflected signal as compared to the diffuse signal. Additionally, as illustrated in the previous paragraph, it also depends on the variation of the reflected signal compared to the diffuse one. The transition between correct and erroneous depth estimate is not gradual but binary. Once the reflection signal variation is stronger than the diffuse signal variation, the wrong disparity is chosen. Before that, the depth estimate is correct. Note that the actual threshold value where this switch occurs also depends on the data term used. Robust Data terms as discussed in the related work can often raise this value but can never completely eliminate it.

6.4.2 Modeling Reflections

The scene parametrization is depicted in Figure 6.5. The world is assumed to be representable on image grid Ω, where each pixel $i \in \Omega$ represents a surface element parametrized by radial distance r_i from the primary (left) camera center, surface normal orientation (θ_i, ϕ_i), diffuse color f_i and additional material parameters $(\mu_i, [\sigma_i])$. Note that though r_i is a scalar, it implicitly corresponds to a 3D point and also a ray direction by a function $\mathbf{x}^v(r_i)$, defined only by the (known) camera parameters in

v. When the superscript v is omitted, I refer to a 3D point in the primary camera system. Similarly, (θ_i, ϕ_i) define the normal $\mathbf{n}(\theta, \phi)$ and r_i and $\mathbf{n}$ together define a plane $\mathbf{p}(r, \mathbf{n})$. Wherever it eases readability, I simply refer to these derived values as $\mathbf{x_i}$, $\mathbf{n_i}$ and $\mathbf{p_i}$ respectively. The color vector f_i is required only for the derivation of th model. With the simplifications that will be made, we will see that f_i can be implicitly recovered from the observed color using the other parameters.

For each pixel I define a vector of unknown parameters

$$\mathbf{s}_i = \{r_i, f_i, \theta_i, \phi_i, \mu_i, \sigma_i\}. \tag{6.1}$$

With bold face capital letters I refer to the set of a single parameter over all pixels, e.g. $\mathbf{R} = \{r_i\}, i \in \Omega$, $\mathbf{S} = \{\mathbf{s}_i\}, i \in \Omega$. Next, I define V as the set of cameras defined by their extrinsic and intrinsic parameters and the mapping

$$\pi^v : \mathbb{R}^3 \to \mathbb{R}^2, v \in V, \tag{6.2}$$

that projects 3D world points into view v. When applied to scalar r_i, define

$$\pi_i^v(r) = \pi^v(\mathbf{x}_i(r)), \tag{6.3}$$

which takes depths at pixel i in the reference view to view v. Finally, let C be a color space and let

$$I^v : \mathbb{R}^2 \to C, v \in V \tag{6.4}$$

map the position on the 2D image plane of camera v to the observed color at this point with bilinear interpolation for non-integer coordinates. The least squares stereo data term can then

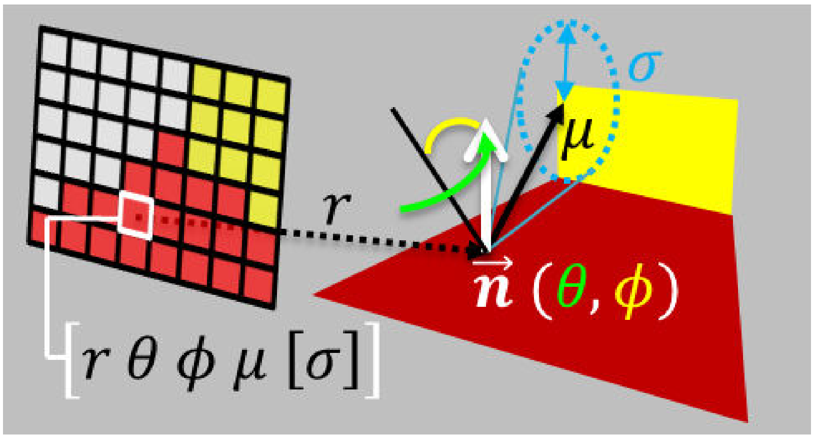


Figure 6.5: Scene Representation. Per pixel, surface geometry is parametrized using the depth r along the pixel ray and the surface normal represented by the Euler angles (θ, ϕ). Materials are represented by a mixing parameter μ and optionally the specular roughness σ. Larger μ corresponds to stronger reflections. Larger σ means more diffuse reflections.

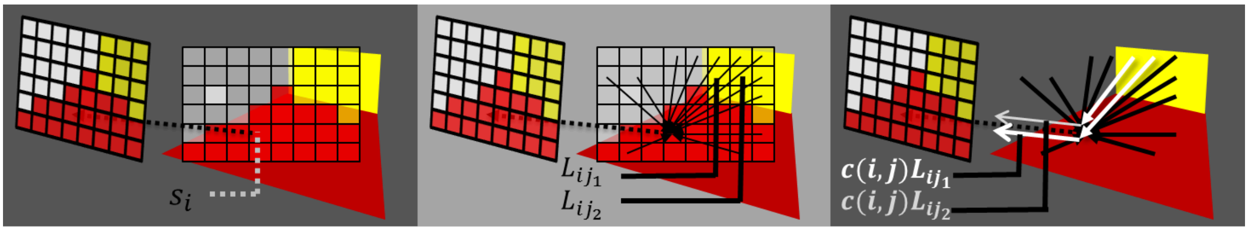


Figure 6.6: Illustration of the Screen Space Version of the Render Equation. The observed color in a pixel is a weighted sum of the diffuse color parameters of all other pixels.

be expressed as the sum of pixel-wise costs

$$LSQ(\mathbf{S}) = \sum_{i \in \Omega} E(\mathbf{s}_i, \mathbf{S}), \tag{6.5}$$

where the pixel-wise cost $E(\mathbf{s}_i, \mathbf{S})$ is defined as

$$E(\mathbf{s}_i, \mathbf{S}) = \sum_{v \in V} ||I^v(\pi_i^v(r_i)) - m(\mathbf{s}_i, \mathbf{S})||_2^2. \tag{6.6}$$

The model function m introduced here computes the observed color of i as a function of the parameters $\mathbf{s}_i$ in i and the set of all other world parameters $\mathbf{S}$. Of course, $\mathbf{s}_i \in \mathbf{S}$, but I want to put an emphasis on the dependency on the parameters of the first surface that is observed.

The observed color from any viewpoint is most generally explained by the rendering equation [92] (cf. Figure 2.1), which, modified to my notation and assuming isotropic light sources is given by (cf. Figure 6.6)

$$m(\mathbf{s_i}, \mathbf{S}) = e_i + \sum_{j \neq i} c(\mathbf{s}_i, \mathbf{s}_j, \mathbf{S}) L_{ij}. \tag{6.7}$$

In essence, this equation states that the color observed at a location (the pixel) from a surface point corresponds to the amount of light e_i that the surface patch emits itself and the fraction $c(\mathbf{s}_i, \mathbf{s}_j)$ of light $L_{ij}(\mathbf{S})$ received from another surface point j that is reflected into the camera pixel. The function c corresponds to a discrete version of the BRDF, which, as a reminder, is a material specific property that governs how surfaces appear under different lighting and viewing angles (cf. Section 2.1.2). Note that in general, the light transported from one surface to another depends on the light that the transmitting surface receives from all other surfaces in the scene etc. There is no analytical solution for the forward problem such that renderers have to employ Monte Carlo or Finite element techniques to compute the full global light transport. w.l.o.g. I assume that the BRDF

c decomposes into a diffuse, viewpoint independent part (i.e. a constant part) and a viewpoint dependent specular part

$$c(i, j) = c_i^{\text{diff}} + c^{\text{spec}}(\mathbf{s}_i, \mathbf{s}_j), \tag{6.8}$$

such that Equation (6.7) can be written as

$$m(\mathbf{s_i}, \mathbf{S}) = e_i + \sum_{\mathbf{s}_j \in \mathbf{S}} c^{\text{diff}} L_{ij} + \sum_{\mathbf{s}_j \in \mathbf{S}} c(\mathbf{s}_i, \mathbf{s}_j, S) L_{ij}. \tag{6.9}$$

Since the amount of light received from the other surfaces is viewpoint independent L_{ij}, I define the diffuse color f_i of the surface point as

$$f_i = e_i + \sum_{\mathbf{s}_j \in \mathbf{S}} c^{\text{diff}} L_{ij}. \tag{6.10}$$

Finally, I make a **single-bounce assumption**: the light received from another surface position only corresponds to its diffuse color. Obviously the model now cannot explain multiple reflections, but this is an approximation required to make the model tractable. Using this approximation, Equation (6.6) can be rewritten as

$$E(\mathbf{s}_i, \mathbf{S}) = \sum_{v \in V} \left\| I^v(\pi^v(r_i)) - \left(f_i + \mu_i \sum_{j \in \Omega} c^{\text{spec}}(\mathbf{s}_i, \mathbf{s}_j) f_j \right) \right\|_2^2. \tag{6.11}$$

The actual model that is now obtained depends on the definition of c^{spec}. In the following, I will show that the standard stereo model is a special case of Equation (6.11) with a diffuse BRDF. Further more, I will present two other models that are of interest and which arise by plugging in other BRDF models. All of these models are illustrated in Figures 6.7, 6.8 and 6.9.

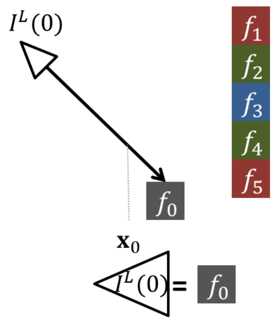


Figure 6.7: DN model. The observed color only depends on the first observed surface. The dependency on surface normals only appears when per-pixel cost is aggregated over a support window as done in Section 6.5.2.

Diffuse World Stereo (DN)[1] For $c^{\text{spec}}(\mathbf{s}_i, \mathbf{s}_j) = 0 \; \forall i, j \in \Omega$, we obtain

$$E(\mathbf{s}_i, \mathbf{S}) = \sum_{v \in V} \| I^v(\pi^v(r_i)) - f_i \|_2^2. \tag{6.12}$$

The solution for $\mathbf{F}$ given two views $V = \{L, R\}$ and depth map $\mathbf{R}$ is

$$f_i^{\text{diffuse}} = \frac{I^L(\pi^L(r_i)) + I^R(\pi^R(r_i))}{2}. \tag{6.13}$$

[1] **D**epth, **N**ormals. The normals are only used in combination with cost aggregation

Replacing f_i in Equation (6.12) with f_i^{diffuse}, we obtain

$$LSQ(\mathbf{R}) = \frac{1}{2}\sum_{i\in\Omega}\left\|I^L(\pi^L(r_i)) - I^R(\pi^R(r_i))\right\|_2^2, \tag{6.14}$$

which coresponds to the standard least squares stereo matching term.

Delta-BRDF Model (DNM)[2] Consider

$$c^{\text{spec}}(\mathbf{s}_i, \mathbf{s}_j) = \begin{cases} \mu_i & \text{if } H(n_i)\mathbf{x}_i(r_i) \times (\mathbf{x}_j(d_j) - \mathbf{x}_i(r_i)) = 0 \\ 0 & \text{otherwise }, \end{cases} \tag{6.15}$$

where $H(v) = I - 2vv'$ is the Householder transform that describes mirror reflection.

This BRDF model corresponds to a perfect mirror reflection which is only weighted by the parameter μ_i. The inner sum in Equation (6.11) reduces to

$$E(\mathbf{s}_i, \mathbf{S}) = \sum_{v\in V}\left\|I^v(\pi^v(r_i)) - \left(f_i + \mu_i f_{\rho(i,v)}\right)\right\|_2^2, \tag{6.16}$$

where the function $\rho(i,v)$ finds the pixel corresponding to the intersecting surface point. In practice ρ has to be implemented by some form of ray tracing. In the next section, I show how this is done efficiently in screen space. Also note how this model is extremely sparse for a fixed choice of surface normals.

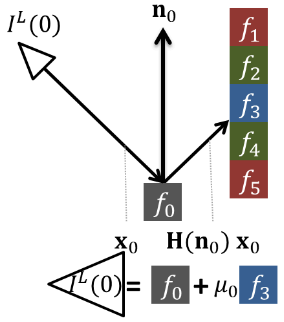

$I^L(0)$
$\mathbf{n}_0$
f_1
f_2
f_3
f_4
f_5
f_0
$\mathbf{x}_0$
$\mathbf{H}(\mathbf{n}_0)\,\mathbf{x}_0$
$I^L(0) = f_0 + \mu_0 f_3$

Figure 6.8: DNM Model. The observed color depends on the diffuse color of the first-bounce f_0 and the diffuse color of the second bounce in mirror direction. The strength of the reflection is governed by parameter μ_i.

Rough Gloss Model (DNMS)[3] Finally, I consider a specular term of the form

$$c^{\text{spec}}(\mathbf{s}_i, \mathbf{s}_j) = \begin{cases} \frac{\mu_i}{M(\mathbf{S},i)} & \text{if } \left\langle \frac{H(n_i)\mathbf{x}_i(r_i)}{||H(n_i)\mathbf{x}_i(r_i)||}, \frac{\mathbf{x}_i(d_j)-\mathbf{x}_i(r_i)}{||\mathbf{x}_i(d_j)-\mathbf{x}_i(r_i)||} \right\rangle < \sigma_i \\ 0 & \text{otherwise }, \end{cases} \tag{6.17}$$

where $M(\mathbf{S}, i)$ is a normalizing factor corresponding to the number of pixels for which the condition is true. This type of BRDF implies a constant value if the angle between mirror reflection direction and direction toward $\mathbf{s}_j$ is smaller than a certain threshold defined by the fourth parameter σ. With this kind of BRDF model, I try to emulate the roughness parameters observed in common BRDF models such as Phong, Gaussian or Ward BRDF

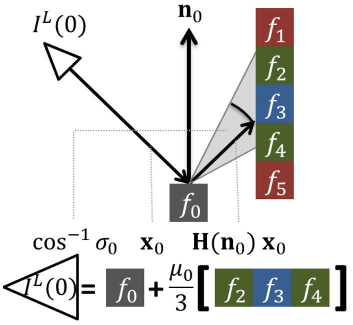

$I^L(0)$
$\mathbf{n}_0$
f_1
f_2
f_3
f_4
f_5
f_0
$\cos^{-1}\sigma_0$
$\mathbf{x}_0$
$\mathbf{H}(\mathbf{n}_0)\,\mathbf{x}_0$
$I^L(0) = f_0 + \frac{\mu_0}{3}[f_2\ f_3\ f_4]$

Figure 6.9: DNMS Model. The observed color depends on the diffuse color of the first-bounce f_0 and on all surface points that are observed under a certain range of angles around the mirror direction. The range of angles taken into consideration depends on the roughness parameter σ_0.

[2] **D**epth, **N**ormals, **M**u
[3] **D**epth, **N**ormals, **M**u, **S**igma

models. The corresponding energy is

$$E(\mathbf{s}_i, \mathbf{S}) = \sum_{v \in V} \left\| I^v(\pi^v(r_i)) - \left(f_i + \frac{\mu_i}{M(\mathbf{S}, i)} \sum_{j \in \tilde{\Omega}(v, \sigma_i, r_i, n_i)} f_j \right) \right\|_2^2 . \tag{6.18}$$

The differences to Equation (6.11) are quite subtle. c^{spec} can be eliminated by reducing the support of the inner sum to those pixels that lie in the valid range. The number of entries in the sum can still get quite large with a larger distance between viewed surface and reflected object, thus making the evaluation of the objective very time-consuming. In the next section a constant time screen space approximation for the computation is presented.

Eliminating first bounce f_i The two Equations (6.16) and (6.18) both have the structure

$$LSQ(\mathbf{S}) = \sum_{i \in \Omega} \sum_{v \in V} \| I^v(\pi^v(r_i)) - (f_i + \mu_i r_i^v(F)) \|_2^2, \tag{6.19}$$

where r_i is the reflection component. Following the same arguments as in the diffuse case, the least square solution for f_i is given by

$$f_i^{\text{diffuse}} = \frac{I^L(\pi^L(r_i)) + I^R(\pi^R(r_i)) - \mu_i \left(r_i^L(F) + r_i^R(F) \right)}{2}, \tag{6.20}$$

yielding the following per pixel matching cost for both Equation (6.16) and (6.18):

$$E(\mathbf{s}_i, \mathbf{S}) = \frac{1}{2} \left\| I^L(\pi^L(r_i)) - I^R(\pi^R(r_i)) - \mu_i \left(r_i^L(F) - r_i^R(F) \right) \right\|_2^2 . \tag{6.21}$$

While f has not been completely eliminated from the cost, it now only appears in the reflection term. In the next section, this is further simplified to compute reflections in screen space.

In the following, I will refer to the two least squares energies derived by applying Equation (6.21) to Equations (6.16) and (6.18) as the DNM and DNMS models respectively. Similarly standard stereo matching with surface orientation (cf. Eq. (6.12)) will be referred to as DN.

Offscreen bounces As written, the model assumes that specular bounces touch parts of the scene that are visible in the image. In a similar vein, some readers may wonder how light sources outside the scene were not mentioned at all. The straight forward answer to this is model tractability. Some form of prior information has to be introduced to estimate anything outside the scene. In the present work, the main interest is in what can be derived only from information available in the image. Therefore, instead of additionally modeling lights and shading as commonly done in shape from shading/intrinsic image research, the diffuse shading and diffuse reflection of lights as well as surface emissivity are just part of the diffuse color f_i. Similarly, if the diffuse color is not limited to lie in the [0, 1], f_i can also model observed light sources. The case when the model is violated is when a specular surface introduces the off screen bounces. As mentioned, handling these is subject to future work.

6.4.3 Model Validation

Before discussing inference strategies, I first present experiments intended to verify that the correct solution also has the lowest least squares energy. Experiments were undertaken on small (32x32) px synthetic images rendered using Blender with known ground truth and the DNMS model.

Random Sampling A varying percentage of ground truth parameters were perturbed with different levels of additive Gaussian noise. The set of feasible reflection pixels were computed by brute force comparison of the spatial relationship between all pairs of pixels. The results are depicted in Figure 6.10. For all but 60 tested realizations, the perturbed parameters have a higher residual energy than the correct solution. These cases occurred at the lowest noise level, where the parameters are very close to the real solution.

Energy Surfaces To further verify the validity, energy surfaces of slices around the ground truth parameters were computed and are visualized in Figure 6.11. The first experiment varied the parameters along the μ - σ plane with all other parameters being ground truth. The energy minimum is close to the ground truth parameter set, but does not coincide with it. As the μ does have the right value of 0.7, this can be explained by the simplified DNMS roughness model as compared to the Beckmann microfacet Model used by the renderer. The second experiment varied the normals around the ground truth normals. With the ground truth normal pointing in direction of the y axis, the perturbed normals were changed using rotating by angles α_x and

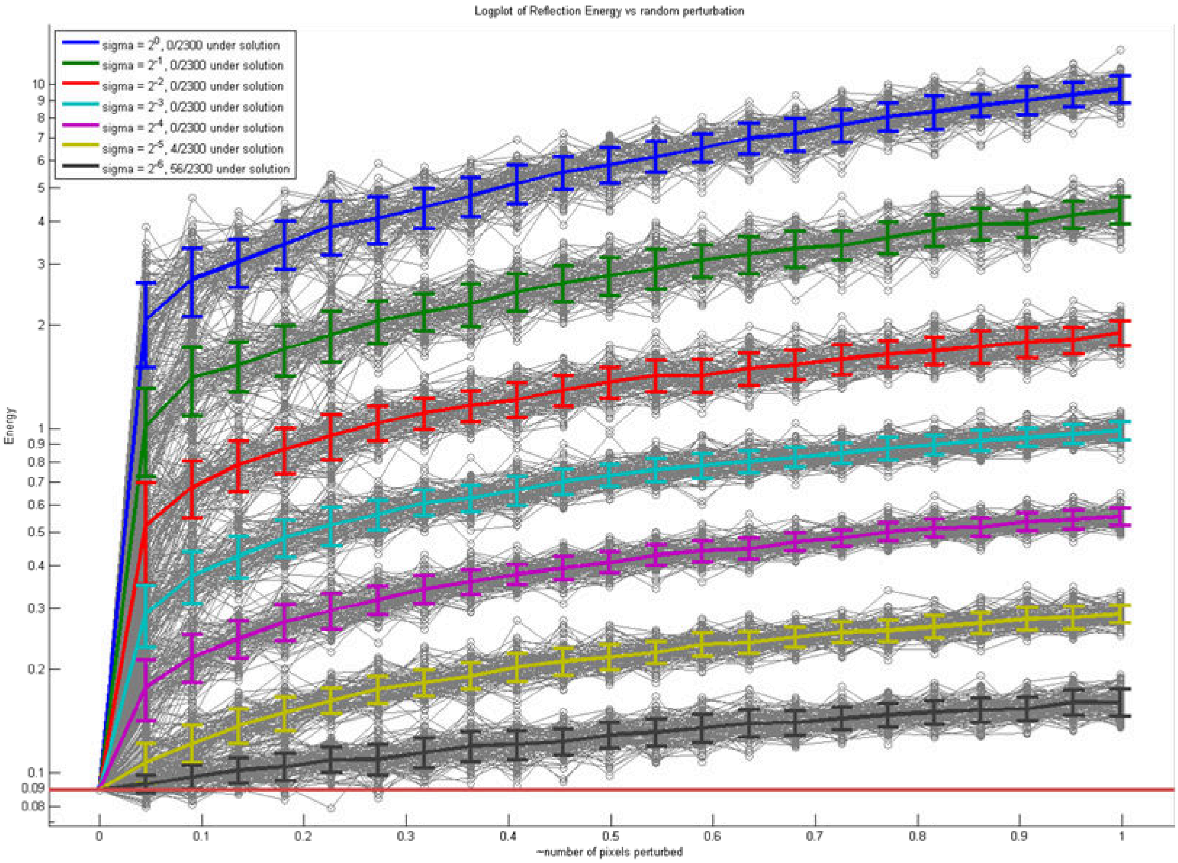


Figure 6.10: Model Verification by Random Sampling for DNMS. The parameters were randomly perturbed around the correct solution. Individual lines correspond to different levels of noise. All measured values are plotted in gray addition to mean and standard deviation of repetitions. The proposed DNMS model has the lowest LSQ residual energy for all but 60/16100 tested realizations. These 60 cases occurred at the smallest noise level, where the parameters were very close to the correct solution.

α_z and around the x and z axes respectively For this case, the ground truth parameter set corresponds to the energy surface minimum. Also, both surfaces seem smooth around the desired solution such that gradient based methods could work if initialized sufficiently close to the right solution.

6.5 Inference

The DNM and DNMS models still have large non-local interactions. This is because the reflected color observed in a certain pixel still depends on the geometry of the whole scene and on the diffuse color of the reflected pixels. Finding the right inference technique was therefore subject to quite an amount of trial and error. Each of the attempts resulted in new insights that to some extent have been discussed in Section 6.4.1 and to some extent motivated the method that I present in Section 6.5.2. The attempts and insights will therefore be presented briefly in the next section first.

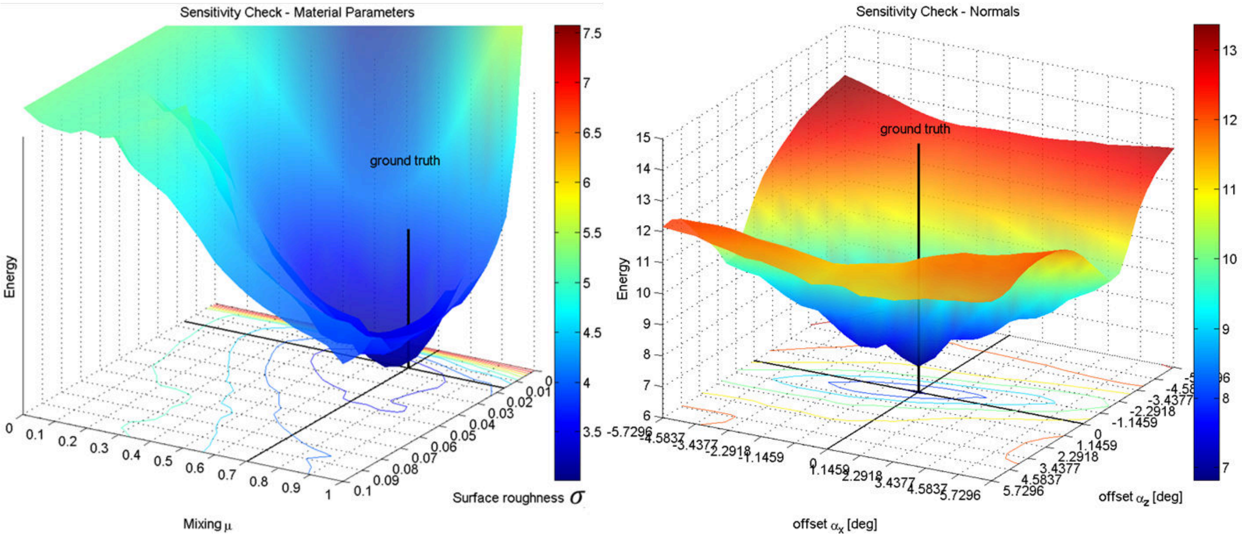


Figure 6.11: Energy surfaces of the DNMS model along parameter slices around the ground truth parameter set. **Left**: Surface along μ- σ slice. **Right**: Surface depending on normal orientation along α_x - α_z .

6.5.1 Lessons Learned from Early Approaches

The DNM and DNMS models have large non local interactions as the reflected color observed in a certain pixel depends on the geometry of the whole scene and the diffuse color of the reflected pixels. Finding the right inference technique was therefore subject to quite an amount of trial and error. Each of the (failed) attempts resulted in new insights that may of interest to some readers and will be presented in the following.

Direct Minimization The method that suggests itself and that was also tried first is to directly solve the energy minimization problem given by the render equation. Since each pixel can potentially interact with any pixel, this requires the computation of derivatives of the per pixel energy with respect to the parameters in all other pixels. Other than being slow for all but the smallest of problems[4], the energy surface does also have many local minima, especially if the diffuse color **F** is left in the optimization. Given an arbitrary depth and material constellation, the nearest local minimum can always be found by setting the diffuse color to be a mix of the left observed image and whatever the current geometry projects to in the right image. The issue of the computation of a large dense Jacobian can be countered to some extent by observing that for a given constellation of geometry, the number of interactions is actually small in the DNM model and also in the DNMS model if only short range interactions are

[4] A 16×16 px color image has a DNMS Jacobian that is 2048×768, which has to be evaluated for each iteration of the continuous minimizer.

taken into account. Therefore, in these cases, the Jacobian is in fact sparse, but the structure of sparsity can change in every iteration, depending on the current set of parameters (e.g. that the reflected mirror reflection point depends on the geometry of the scene). Another approach we therefore tried was to compute the pixel interactions in an outer loop and then to apply continuous minimization on this sparse model in an inner loop. The problem observed here is that, by fixing the structure of interactions between pixels, it is no longer possible to compute derivatives of the reflected color with respect to the change in normals $\partial_{n_i} r_i$. As it turns out, this is quite essential towards solving the inference problem. On a similar note, it should be observed that the DNM model, in essence, makes it impossible to calculate the $\partial_{n_i} r_i$ without any further approximations.

Scale Space Approaches Another approach considered for handling the large parameter space was to utilize a scale space approach commonly employed to tackle spurious local minima as well as the large computational burden when applying continuous/variational techniques to reconstruction problems. Such algorithms operate on a scale space pyramid constructed from the original observed image(s). For the case of stereo matching and in its simplest form the algorithm works as follows: N filtered images $I_k^V, k = 1, ..., N$ are created by filtering the observed images I^V using a Gaussian filter with standard deviation σ_b/α^k, where σ_b is the Gaussian standard deviation at the largest scale and α is the ratio between scales. Then, the idea is to iteratively apply the stereo algorithm on successively smaller scales starting at the largest one. The smoothing process eliminates high frequency variations of the image and therefore also the high frequency variations of the energy (and thus the local minima). By successively applying the method on decreasing scales, the location of the local minimum can then be refined if the previous iteration decreases the distance to the right minimum. This kind of coarse-to-fine inference strategy makes the basic assumption that the optimal parameter set is consistent over different scales. For reflective surfaces this is frequently not the case. The majority of surfaces have some form of micro facet roughness causing diffuse specular reflection. Since diffuse reflection components bear some similarities to a low-pass or smooth version of the perfect mirror reflection, reflection components and diffuse signal from the first bounce surface may frequently appear at different scales. At a coarse scale, there may therefore only be evidence for the erroneous depth given by the reflection component and no evidence of the actual depth.

Planar Proposals Finally, an ad-hoc approach was attempted which used planar proposals sampled from the diffuse stereo result. The motivation behind such a strategy was that errors caused by reflections are binary in nature (i.e. if the diffuse signal is strong enough, the estimated geometry is correct) and that erroneous regions often lie on the same surface next to regions with the correct geometry. After computing the depth map using a diffuse stereo approach, the per pixel normals were estimated using a $3 \times 3\,\mathrm{px}$ neighborhood of the depth map around each pixel. From this depth and normal map, planar proposals with random material properties were created. These planar proposals were used to estimate the direction of reflection and the first bounce color. The second bounce color was then computed using the diffuse stereo geometry. This proposal was then fused with the original diffuse world solution using fusion moves [108]. While this approach did work for small images, it failed on larger images. The first issue was that the normals derived from the diffuse stereo result were not accurate enough for sampling purposes, especially on larger images where small errors in normals result in large (in number of pixels displacement) changes of the reflected image. The other issue was that such a method could not tackle curved surfaces very well.

6.5.2 Continous Data-Driven PatchMatch

Summarizing the findings presented in the previous section, direct continuous minimization of the energy is slow and does not yield any useful results as the optimization converges to local minima. Variational techniques need a good initial guess or a scale space approach. Cues for the actual surface and the reflected surface appear on different scales, thus violating the basic assumption of scale space approaches that the estimated depth is consistent over all scales. Planar proposals sampled from the diffuse stereo result showed promising results on small images containing a large planar surface. However, they failed on curved surfaces and on larger images as normals obtained from the diffuse stereo result did not have the required accuracy on larger images.

Yet, the approach using planar proposals sampled from the diffuse stereo result seemed to be most promising, but needed some modifications:

- The planar proposals chosen have to not only be sampled from the diffuse stereo solution but also around it,
- the planar proposals have to be defined locally to account for curved surfaces,
- the estimated normals of diffuse stereo need to be more accurate to cater for sampling and finally,
- since the reflected image is very sensitive to the change in normals, we require a method for computing derivatives of the reflected image with respect to the surface normals.

PatchMatch Stereo [19] offers the first two properties and therefore lent itself as the framework for inference. In the following, I show that by making some further simplifications to the model and some extensions to the framework, both DNM and DNMS models can be solved using stereo Patchmatch. The basic strategy is to first solve for standard diffuse stereo to obtain an initial guess for geometry. To achieve normal estimation of sufficient accuracy, PatchMatch with continuous refinement is required. This novel extension to PatchMatch inference is described below. Using this initial guess, again two iterations of continuous PatchMatch are applied using the DNM or DNMS models respectively to obtain the final result.

PatchMatch Stereo revisited

PatchMatch stereo operates on an extended cost volume

$$C : \omega \times \mathbb{R}^N \to \mathbb{R}, (i, \mathbf{s}_i) \to E^{pm}(\mathbf{s}_i), \tag{6.22}$$

which outputs the cost for assigning parameter $\mathbf{s}_i$ to pixel location i. E^{pm} is usually defined as a basic pixel cost $E(s_i)$ aggregated over a support neighborhood $N^{pm}(i)$ around i, with

$$E^{pm}(s_i) = \sum_{j \in N^{pm}(i)} w\left(I^L(i), I^L(j)\right) E\left(\tau(j, s_i)\right). \tag{6.23}$$

Here, $w(I^L(i), I^L(j))$ is an optional weighting term that can either be constant 1 or an color adaptive support weight

$$w_{ij} = w\left(I^L(i), I^L(j)\right) = \exp(\gamma^{-1}|(I^L(i) - I^L(j))|). \tag{6.24}$$

The mapping τ transforms the $\mathbf{s}_i$, which is represented according to pixel i into a representation according to pixel j. For standard fronto-parallel stereo where disparities d_i are estimated ($\mathbf{s}_i = d_i$), the mapping is

$$\tau^D : (j, d_i) \rightarrow d_i. \tag{6.25}$$

In [19] it is assumed that the patch geometry can be described by a slanted plane, therefore we get

$$\tau^{DN} : (j, \{r_i, n_i\}) \rightarrow \{||\mathbf{x}_j \cap \mathbf{p}_i||, \mathbf{n}_i\}. \tag{6.26}$$

The $\cap$ symbol denotes the intersection of the direction given the pixel ray $\mathbf{x}_j$ and the plane $\mathbf{p}_i$ defined by $(r_i, \mathbf{n}_i)$. This maps the normals as they are, but transforms the depth such that it belongs to the same plane as the surface described by $\mathbf{s}_i$ in pixel i.

The algorithm then operates as follows: for initialization, the $\mathbf{s}_i$ are drawn randomly from the feasible set of parameters. Then two steps are alternated for each pixel and each pixel is traversed in some order.

In the **propagation** step, the current parameter set in i is replaced by

$$\mathbf{s}_i^{new} = \underset{j \in N(i)}{\arg\min}\, E(i, \tau(i, \mathbf{s}_j)), \tag{6.27}$$

where $N(i)$ describes some neighborhood of i. Note that this is the same τ as defined above, but instead of applying the same s_i to several neighboring pixels, we are now choosing the neighboring $\mathbf{s}_j$ that gives the smallest cost when applied to the current pixel i.

In **random refinement** the current parameter set in i is then refined by drawing random parameters around the current parameter according to some probability distribution $\mathcal{D}(\mathbf{s}_i, \alpha)$ centered around $\mathbf{s}_i$ with additional parameter α that usually corresponds to the variance of the distribution

$$s_i^{new} = \underset{\mathbf{s} \sim \mathcal{D}(\mathbf{s}_i, \alpha)}{\arg\min}\, E(i, \tau(i, \mathbf{s})). \tag{6.28}$$

In stereo PatchMatch this $\mathcal{D}$ corresponds to a double exponential distribution. [5] An intuition and a proof of why this technique works are given in [11]. To sum it up, the method works well if the scene consists of large homogeneous areas with the same or

[5] The sampling strategy employed additionally stratifies the samples into quantile brackets.

slowly varying parameter sets. This is to some extent true for general depth maps and more so for materials, as natural scenes often only consist of a few different materials.

Proposed PatchMatch Variant

The proposed PatchMatch variant makes three modifications to the original PatchMatch implementation. First, the random refinement step is extended to also do gradient based refinement. This step significantly improves the quality of the estimated geometry even for diffuse PM. Next, the per pixel cost is modified in such a way that it does not depend on the parameters in the other pixels. This enables the application of continuous PM also to the DNM and DNMS models. Finally, new data driven sampling routines are employed for the random refinement step.

Continuous Refinement If the pixel-wise cost is defined in such a way that the Jacobian $J_E(\mathbf{s}_i)$ with respect to $\mathbf{s}_i$ can be computed, it is possible to find the local cost minimum using gradient descent or trust region solvers. For the DN model this is evident if linear or higher order spline interpolation between pixels is employed. For DNM and DNMS this becomes a bit more challenging since the evaluation of the cost requires a ray-tracing step. It is also important to be able to compute the derivatives of the reflected color with respect to the change of normal orientation. In practice, the continuous part of the optimization is implemented using the Ceres-Solver [4] library, which computes exact derivatives using automatic differentiation techniques [150]

Screen Space Reflection Computation For continuous optimization of DNM, I further approximate the model by assuming that after the first bounce the scene can be described as a plane parametrized by the orientation and distance of the reflected midpoint. The reflected color is then obtained by projecting the intersection of the reflected ray and this plane into the left camera image. If $\mathbf{s}_j$ is the reflected point, the reflected color is computed as

$$r_i^v = I^v\left(\pi^v\left(\mathbf{p}_j \cap H(\mathbf{n_i})\mathbf{x}_i^v\right)\right). \tag{6.29}$$

There is a closed form term for each of the components, which is why derivatives can be easily computed. Also note that I have approximated the diffuse color with the observed color I^v in the

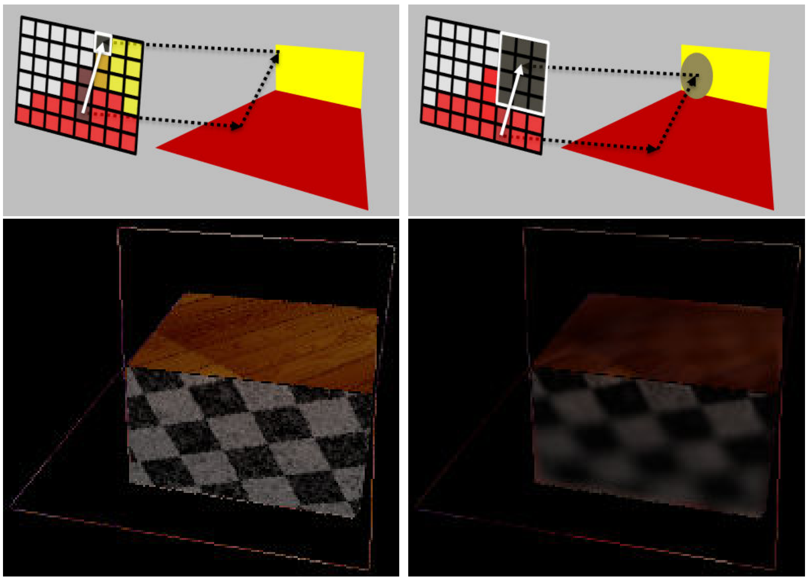


Figure 6.12: **Top**: Scene representation and screen space reflection computation for DNM (left) and DNMS (right) models. **Bottom**: Estimated reflection components for DNM (left pair) and DNMS (right pair). Note that the μ was not multiplied onto the results here such that the top plane reflections are not suppressed. For the DNMS calculations the expected distance based smoothing of the reflection component can be seen.

image. If the reflected point is subject to strong specular color variations, this approximation does not hold anymore.

Implementation-wise the set of parameters $\mathbf{s}_i$ is extended to contain the reflected plane $rp_i = p_j$. p_j is computed using the ray tracing method described below each time the τ mapping is invoked.

To efficiently compute the point of intersection, I borrow screen space rendering techniques known in computer graphics. This is illustrated in Figure 6.12. The reflected ray projected onto the image corresponds to a line direction, thus reducing the number of pixels that have to be tested against. Line search on a grid is done using the Bresenham algorithm [28].

For the DNMS model, the contributions of many pixels have to be taken into account. In the worst case, the reflected color corresponds to the mean color of all surrounding pixels. Evaluating this cost can therefore consume a large amount of time. We simplify this in two steps depicted in Figure 6.12, yielding a constant computational overhead, irrespective of the area of integration. The projection of the reflection cone and the plane of reflection corresponds to an ellipsoid region of integration. The support-region is approximated using a rectangular shape. Integration of rectangular patches can be done efficiently with integral images [40, 182] using four operations irrespective of support size and thus yielding constant time computation irrespective of the choice of sigma and the distance of the reflected point.

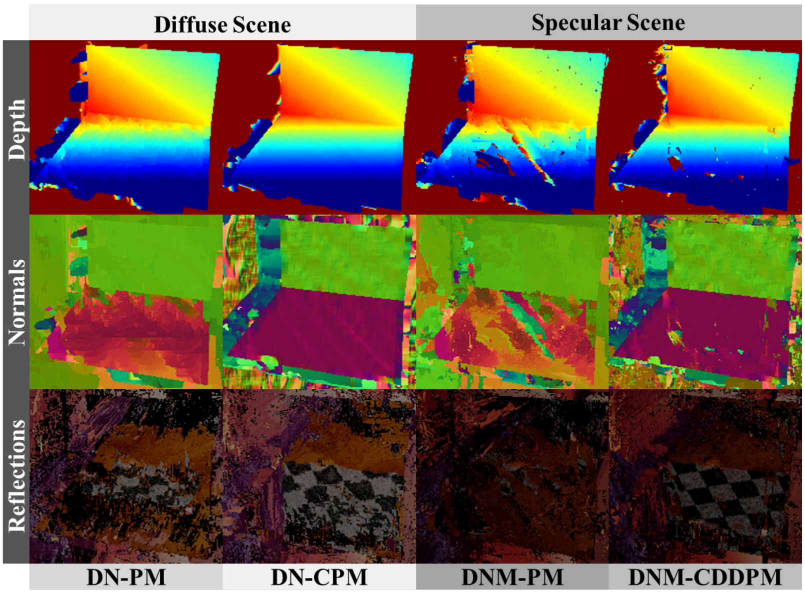


Figure 6.13: Quality of Normals. **Top to bottom**: Jet-coded depth map (4-6 m), RGB-coded normals (components of normals mapped to r,g and b channel) and computed reflections. **From left to right**: a) and b) results of standard PatchMatch (PM) and continuous PatchMatch (CPM) on a diffuse scene. Little difference can be seen in the depth maps. Yet, the normals and reflected images reveal a more accurate geometry estimation. c) and d) result of the DNM model using pm and continuous data driven CPM. Using standard PM estimated normals shows large errors corresponding to the remaining artifacts in the depth image (the results still improve on the DN-PM result on this image (cf. 6.17, first row)). Overall CPM yields better results and in fact improves the quality of normals in the areas of reflection as compared to case b).

Data Driven Sampling Finally during random refinement of the DNM/S models, we replace $\mathcal{D}(\sigma, \theta)$ with a screen space sampler. Given current reflected position $\mathbf{s}_j$, the sampler uniformly samples neighboring pixels as candidate reflection points $\tilde{\mathbf{s}}_j$. The orientation parameters are then computed such that they satisfy $\tilde{\mathbf{s}}_j$ to be the primary point of reflection. This sampling is done additionally to the standard exponential sampling of orientation to allow for searching the proximity of $\tilde{\mathbf{s}}_j$ more closely.

6.5.3 Implementation Details

I present some implementation details and formulas to ease reimplementation. It should be noted that I did not explicitly compute the Jacobian for differentiation, but rather used the automatic differentiation functionality of the Ceres-Solver library [4].

Continuous Images Interpolation was used to obtain an intensity value I^v at any arbitrary position and to compute derivatives with respect to the location. The simplest form, which was used in the experiments reported, was bilinear interpolation

$$I(x, y) = \begin{bmatrix} \lceil x \rceil - x \\ x - \lfloor x \rfloor \end{bmatrix}^T \begin{bmatrix} I(\lfloor x \rfloor, \lfloor y \rfloor) & I(\lfloor x \rfloor, \lceil y \rceil) \\ I(\lceil x \rceil, \lfloor y \rfloor) & I(\lceil x \rceil, \lceil y \rceil) \end{bmatrix} \begin{bmatrix} \lceil y \rceil - y \\ y - \lfloor y \rfloor \end{bmatrix}, \tag{6.30}$$

where $\lfloor\rfloor$ and $\lceil\rceil$ denote floor and ceiling operations. I used the SplineImageView class in the VIGRA computer vision library [103] that allows spline interpolation of arbitrary spline order and also delivers the corresponding derivatives. Initial experience with different spline orders suggest that using cubic spline interpolation instead of linear as commonly used may in fact further improve the accuracy of continuous techniques.

DNM Implementation wise, the set of parameters s_i were augmented by the parameters of the two planes $\mathbf{p}_i^L$ and $\mathbf{p}_i^R$ that the reflected ray intersects. The planes are left constant during continuous optimization and are only changed when the τ function is invoked during propagation (which triggers the screen space ray-tracing step). Each plane is characterized by its normal[6] and offset, i.e. $\mathbf{p}_i^L = (\mathbf{n}_i^L, \beta_i^L)$. If j contains the parameters of the mirror reflection point for one of the camera views (e.g. L), then

$$\mathbf{n}_i^L = \mathbf{n}_j \tag{6.31}$$

and

$$\beta_i = \mathbf{x}_j \cdot \mathbf{n}_j. \tag{6.32}$$

The normal $\mathbf{n}_i$ can be computed from (ϕ_i, θ_i) using standard polar to Cartesian coordinate transform. The screen space positions of the point of reflection i_r^L and i_r^R are found as follows: With $(p_x, p_y) = i$, I refer to the coordinate components of image point i. f is the focal length of the cameras and b the baseline. For the sake of readability, the principal point is assumed to lie in the origin. Let the ray direction of a surface point to each camera be given by

$$\mathbf{x}_i^L = r_i \cdot \frac{(p_x, p_x, f)}{||(p_x, p_x, f)||}, \tag{6.33}$$

and

$$\mathbf{x}_i^R = r_i \cdot \frac{(p_x, p_x, f)}{||(p_x, p_x, f)||} - (b, 0, 0). \tag{6.34}$$

The line of mirror reflection mirror^v is parametrized as

$$\text{mirror}^v(\lambda) = \mathbf{x}_i + \lambda H(\mathbf{n}_i)\mathbf{x}_i^v. \tag{6.35}$$

The point of intersection between this line and the plane of first reflection p_i^v is given by mirror$^v(\lambda^v)$ with

$$\lambda^v = \frac{\beta_i - \mathbf{x}_i \cdot \mathbf{n}_i^v}{H(\mathbf{n}_i)\mathbf{x}_i^v \cdot \mathbf{n}_i^v}. \tag{6.36}$$

[6] $\mathbf{n}_i^L$ should not be confused with $n\mathbf{n}_i$, which is the normal of the first bounce surface.

The point of intersection $\mathrm{mirror}^v(\lambda^v)$ can then be projected back into the left (or right) camera view to obtain the screen space point of reflection i_r^v.

DNMS For the DNMS model, the point of intersection i_r^v in screen space is computed in the same way as for the DNM model. To then retrieve the reflected color from the integral image, the area of integration is additionally required. The integral image itself is computed from the observed images in a preprocessing step. The same spline interpolation technique was applied on the integral image in order to be able to compute smooth integrals that can be differentiated with respect to the model parameters. The box width of the integration is obtained as follows: Let m_z be the z component of the mirror reflection $\mathrm{mirror}^v(\lambda^v)$. The half-width of the integration domain w in pixels then obtained as

$$w = 0.5 + f \cdot \frac{\sigma_i * ||\,\mathrm{mirror}^v(\lambda^v) - x_i||}{m_z}. \tag{6.37}$$

While this is a really coarse approximation to the actual DNMS model, it does the job quite well in practice. The 0.5 constant term is to ensure that the integration takes place over at least one pixel.

6.6 Experiments and Results

In the following, I refer to standard PatchMatch with PM and, similarly, to continuous (data driven) PatchMatch with CPM and CDDPM. Additionally, I prefix the inference method with the model that is to be inferred. DN-PM, DNM-PM, DNMS-PM therefore refer to standard PatchMatch inference using the DN, DNM and DNMS models respectively. Note that only DNM-X and DNMS-X require the screen space reflection computation described earlier. The algorithms utilized a patch window of size 13 px and an exponential color based adaptive support weight (ASW) [19] with parameter 0.08 for images normalized between 0 and 1. The DNM and DNMS models are more sensitive to the choice of color-based ASW since strong reflected edges that give the primary cues for estimating the material properties also cause a strong down-weighting of pixels. The scenes used in the following experiments were modeled in Blender and rendered using the Blender-Cycles renderer that approximates the global illumination. This allows for ground truth evaluation and verification of the reconstructed parameters.

Quality of Normals In the past sections, I often stressed the importance of accurate normals for reflection handling and calculation. The continuous data driven PatchMatch approach was motivated by this goal. To illustrate the effect of normal estimation on handling reflections, I ran PM and CPM/CDDPM on a fully diffuse scene and a scene containing a specular surface. The results are illustrated in Figure 6.13. For the diffuse scene (left half), the reconstructed depth maps are nearly identical using DN-PM and DN-CPM (small deviations can be observed in detail though). Yet, the normal map reveals large differences. The computed mirror reflection using these normals also confirms these findings. For the specular scene (right half), I compare two iterations of DNM-PM after two iterations of DN-PM with 2 iterations of DNM-CDDPM after two iterations of DN-CPM. The differences here are more striking both in depth and normals. Notice the (erroneous) low frequency normal error of the lower surface in DN-CPM, which is no longer present in DNM-CDDPM wherever the surface reflected something else in the scene. This is a strong indicator that modeling reflection not only can correct errors due to reflections, it actually can aid in more accurate geometry estimation. The improved microstructure of the lower surface is further evidenced by the quality of the computed reflections. Finally, some artifacts can still be observed in the DNM result. These are due to reflections of occlusion boundaries that have a similar effect as the ones normal occlusions have in standard stereo. This is not a shortcoming of the model per se, but a result of the simplifications made to compute the reflected color. Possible solutions will be discussed in the conclusion.

DNM/S Model Verification I compare DN-PM with DN-CPM, DNM-CDDPM and DNMS-CDDPM for 11 different scenes of varying curvature of the specular surface. The ground truth BRDF parameters of the specular surface are constant over the whole surface. For the evaluation of the DNM model, for each scene the μ parameter for the lower surface was varied between 0.0 and 0.4. The latter corresponds to a peak diffuse signal to reflection ratio of over 0.6 in this scene. Similarly I report results for the DNMS model with $\mu = 0.25$ and $\sigma = 0.01, 0.02, 0.04$ and 0.1.

Visualization of the results can be found in Figures 6.16 - 6.19 for the DNM model and Figures 6.20 - 6.23 for the DNMS model. All images displayed are a result of two iterations of DNM/S-CDDPM over two iterations of DN-CPM as well as the result of DN-CPM. Further number of iterations did not change the results much, thus suggesting convergence of the methods. The test scene names indicate the curvature of the specular surface in z and x direction, with the prefix 'p' indicating a surface curved towards the sky and 'n' indicating the opposite.

Additionally, Figure 6.14 and Figure 6.15 quantify the results for all tested parameters and scenes for the DNM and DNMS models respectively. In these plots, I report the decrease in the number of 'bad' pixels between results using DN-CPM and results using DNM/S-CDDPM for each of the parameters:

- For the depth, I report the decrease in number of pixels belonging to the foreground object whose ground-truth depth error exceeds 1 cm.
- Similarly, for the surface orientation the decrease in number of pixels belonging to the foreground object is reported, if the ground truth angular error of normals exceeds 5°.
- For μ and σ, I report the decrease in number of pixels that exceed the ground truth by 0.05 and 0.009 respectively. Here, the region of interest was chosen to be the region in the image that contains reflections. This is because as a purely local data term, any value of μ and σ yields the same least square energy in areas that do not reflect anything. While for some of the DNM ($\mu = 0.2$ and 0.4) experiments I set all parameters 0 for pixels that weren't reflecting anything in a post-processing step, this was not done for the DNMS experiments.

The choice of performance metrics does not affect the ranking between methods (e.g. mean squared error, median error etc). The metrics mostly correspond to the 3D-space version of the bad-pixel metric commonly used in Middlebury evaluations [159]. They were chosen as they are best suited for the multi-modal, heavy-tailed error distributions that are caused by reflections. Summarizing, DN-CPM consistently decreases the GT error over DN-PM, and DNM/S-CDDPM consistently further decreases the error. The scenes where the relative decrease is low, correspond to the situation where the actual area reflecting something is relatively small (e.g. 0-n45). The remaining

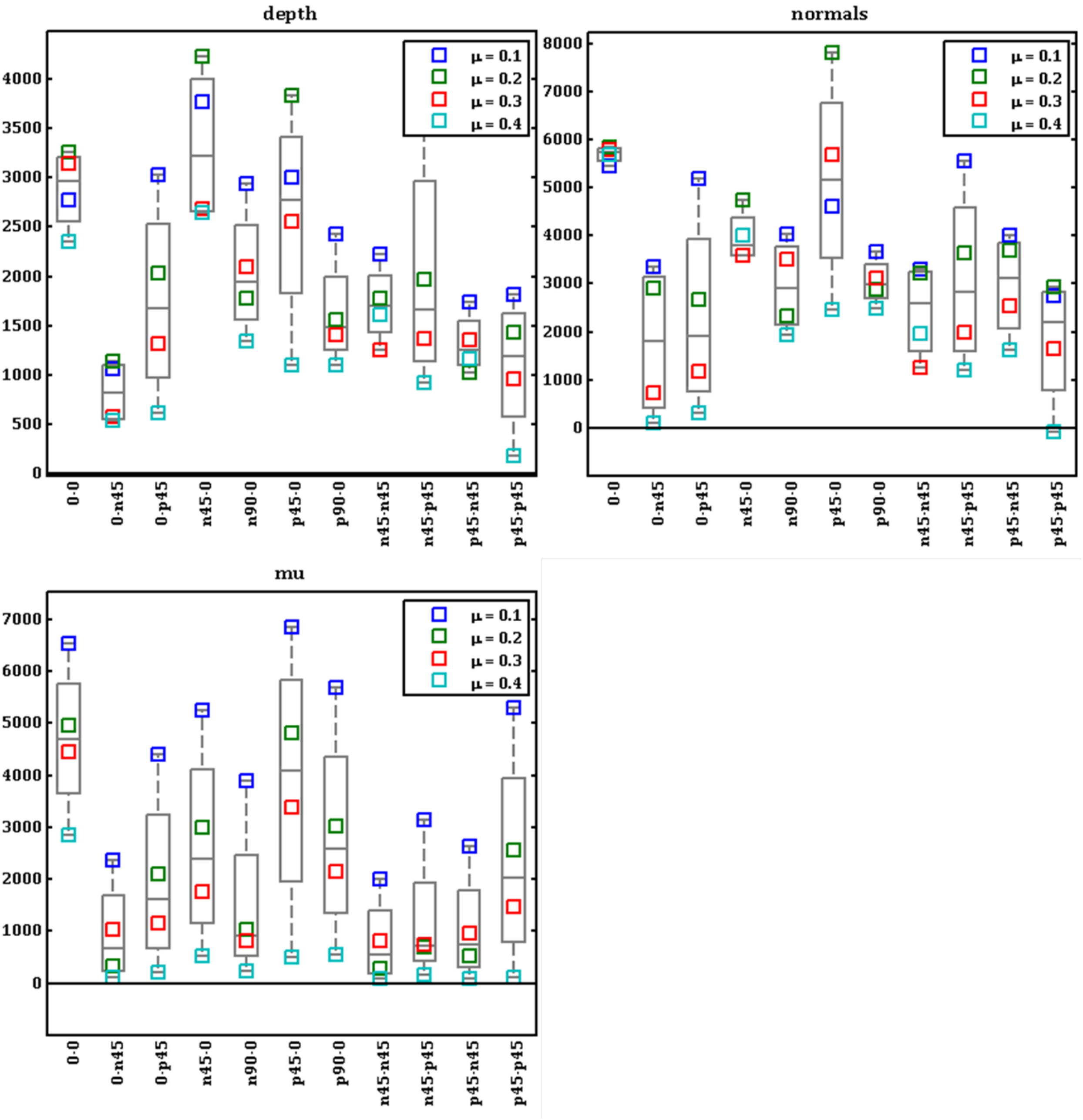


Figure 6.14: Summary of Model Improvement for the DNM Model over the DN Model. These box plots depict the improvement (decrease) of bad pixels as defined in Section 6.6. Larger values mean a lot of improvement, small values mean little improvement. Values under 0 indicate a deterioration. Results for all four tested μ are plotted vertically for each scene. The box plot in gray indicates median and quartile values over all tested parameters for a single scene. We observe consistent improvements of the reconstructed parameters for all three parameters. N.B. for μ this is evident as the DN model does not estimate any material parameters. The overall trend is a deterioration of results for stronger specular components. The scenes with the curved surface in z direction (towards the observer) and the scenes with curvature in two directions are the most difficult to solve while curvature in x direction is easier. Note that this plot only depicts the decrease of the error. For scenes such as 0-n45, where the diffuse result is good already, small values will be observed.

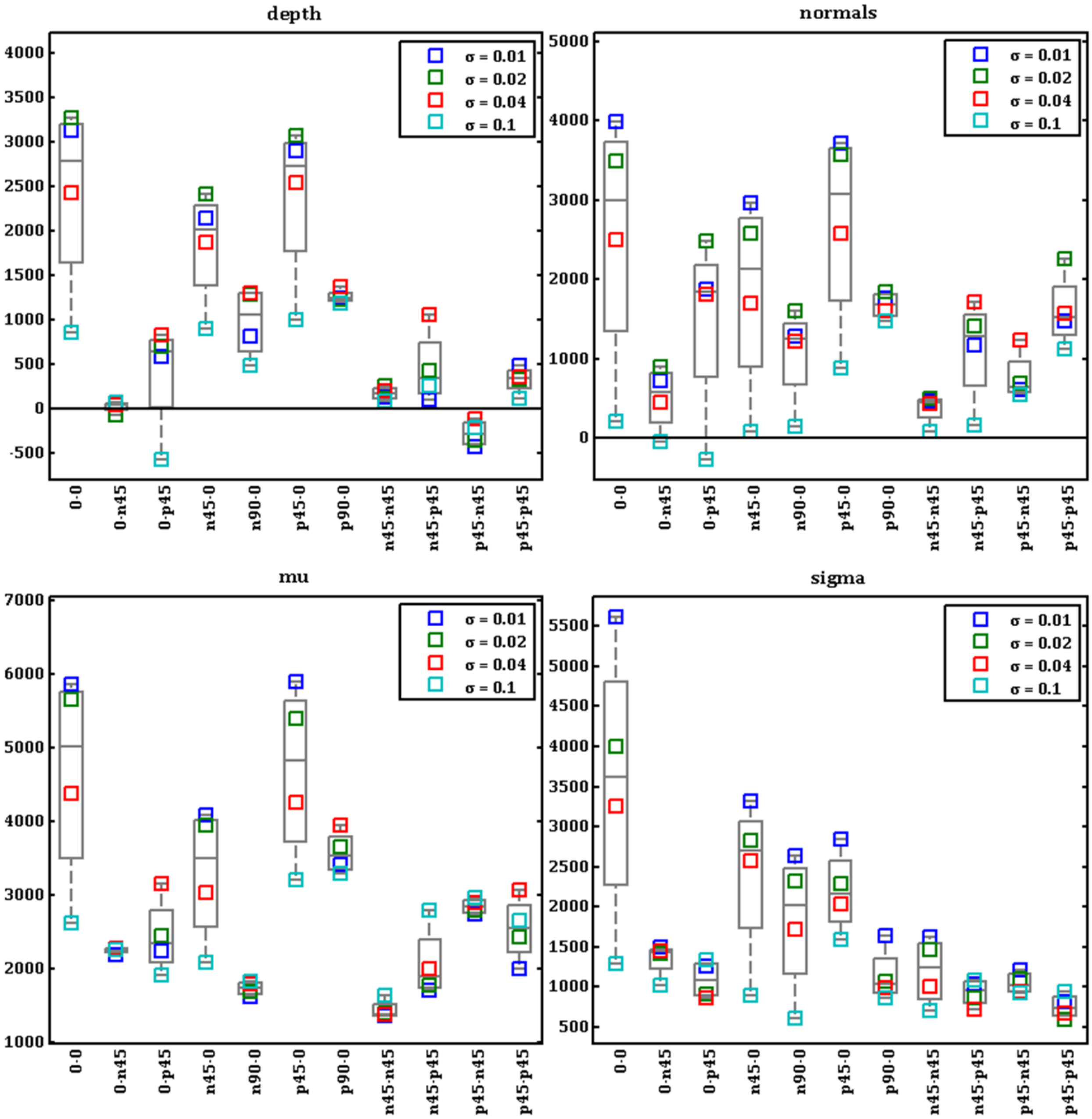


Figure 6.15: Summary of Model Improvement for the DNMS Model. Same observation of overall improvement as above, slightly less pronounced due to the additional parameter. The overall trend is towards a smaller effect for larger σ consistent with large σ corresponding to more diffuse surfaces. Again note that this plot indicates the decrease in error. A low value indicates that the DN and DNM result were close to each other.

artifacts often correspond to the already mentioned reflections of depth edges. Also consistent with the findings above are the normals that are improved upon in areas that reflect other parts of the scene. The proposed method is able to improve the geometry and recover meaningful parameters over a wide range of different surface curvatures. The most difficult situation happens to be a convex surface oriented towards the camera as lot of reflected rays bounce back into the direction of the camera. For larger values of μ, the proposed inference strategy starts to fail, while for larger values of σ the scene becomes indistinguishable from a completely diffuse scene such that the DN model does not produce artifacts. I investigated whether the failure is due to inference strategy or due to model violations. It turns out that the former is the case as the algorithm does not diverge if initialized with the ground truth solution. The main reason for the failing is that the assumption no longer holds true that some parts of the DN-CPM model can be used as an initial guess for normals. The reflected surface resembles a proper mirror more and more and most of the area is erroneous. Again, I verified that this is not an issue with the proposed model as the ground truth solution still has the lowest energy.

6.7 Summary and Outlook

6.7.1 Summary

The work addressed the matter of specular reflections which violate the diffuse world model commonly used for stereo matching. By including the second order terms of the image formation model governed by the render equation, I derived two data terms that are capable of explaining specular reflections. Finally, I showed that the inference of the resulting optimization problem is possible using CDDPM. In consequence, it was possible to estimate depth, normal orientation and material parameters in each pixel. Ground truth evaluation on synthetic datasets shows consistent improvement of estimated parameters and also indicates that by harnessing reflection as opposed to suppressing it, as commonly done in literature, it is possible to estimate geometry with a higher accuracy. The work presented opens up many questions that need to be addressed in the future.

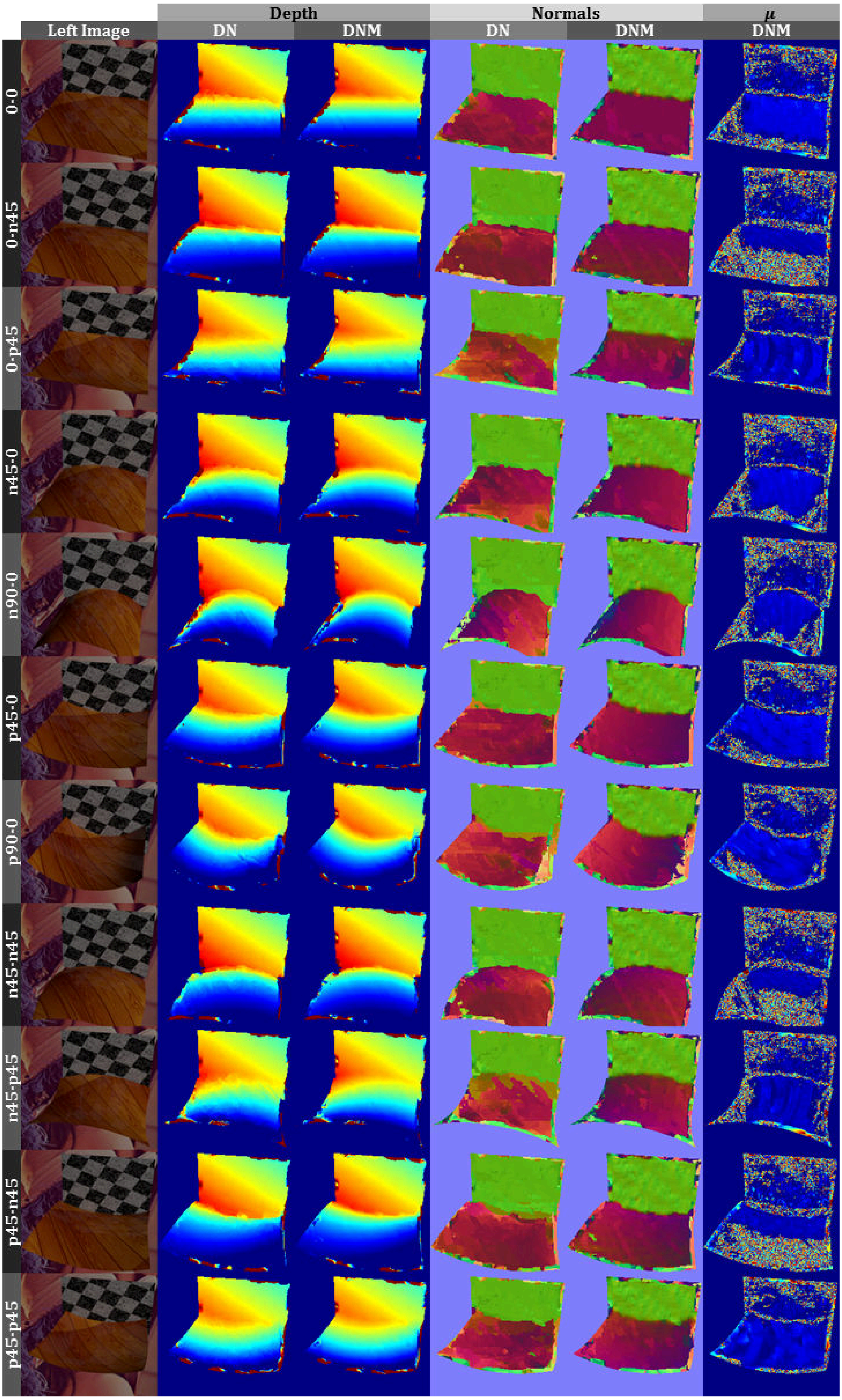


Figure 6.16: All Results for DNM, μ = 0.1. Color coding as in Figure 6.13. μ image in range 0-1. The diffuse component is much stronger than the specular one such that the depth estimates by DN and DNM are nearly identical. Yet, the estimated normals are more accurate, suggesting a more accurately estimated micro structure. Also note the almos perfect material estimation in areas that reflect something. The noise in the material estimates are due to the PM sampling, since surface patches that does not reflect anything (reflect black space) can have any parameter μ .

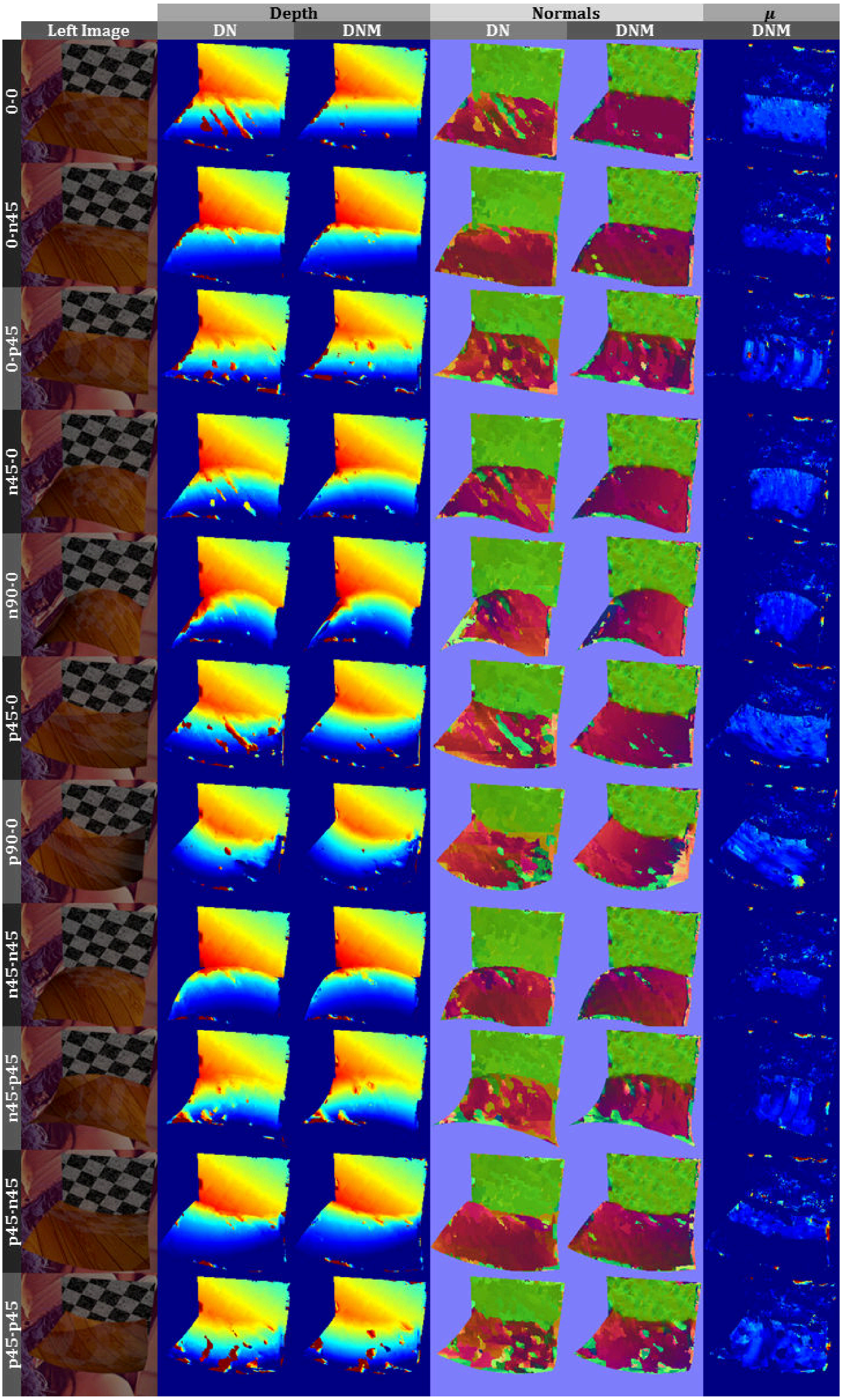


Figure 6.17: All Results for DNM, $\mu = 0.2$. Color coding as in Figure 6.16. For a larger specular term, the reflection signal is stronger. Reflected edges lead to spurious local minima in the diffuse matching cost. By using the proposed algorithm, it is still possible to eliminate many of these errors, improve normals and estimate material. Surfaces curved in x direction seem to be easier to handle than surfaces curved forward (z) direction.

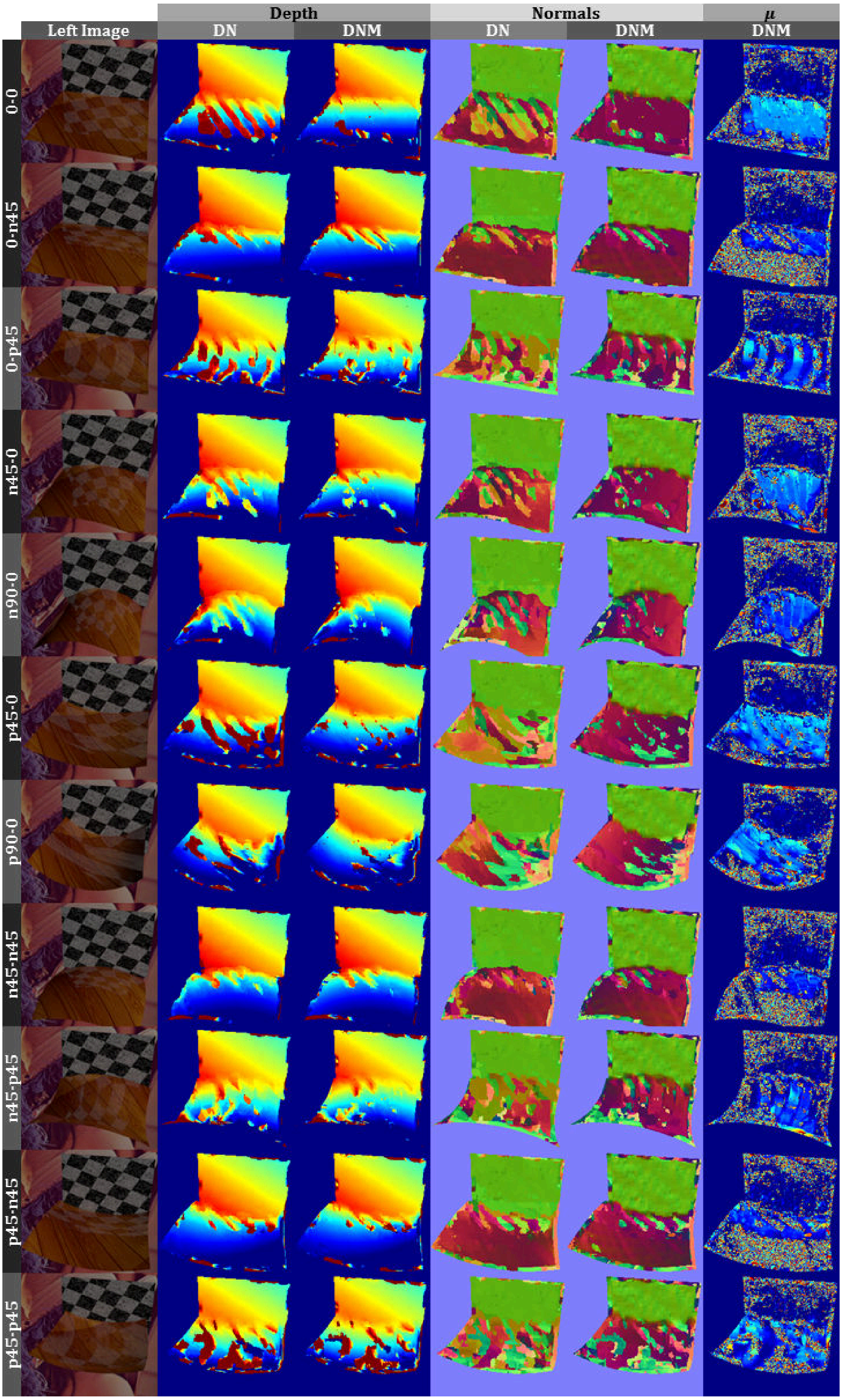


Figure 6.18: All Results for DNM, $\mu = 0.3$. Color coding as in Figure 6.16. The results are mostly similar to the results presented in Figure 6.17 with few more remaining artifacts.

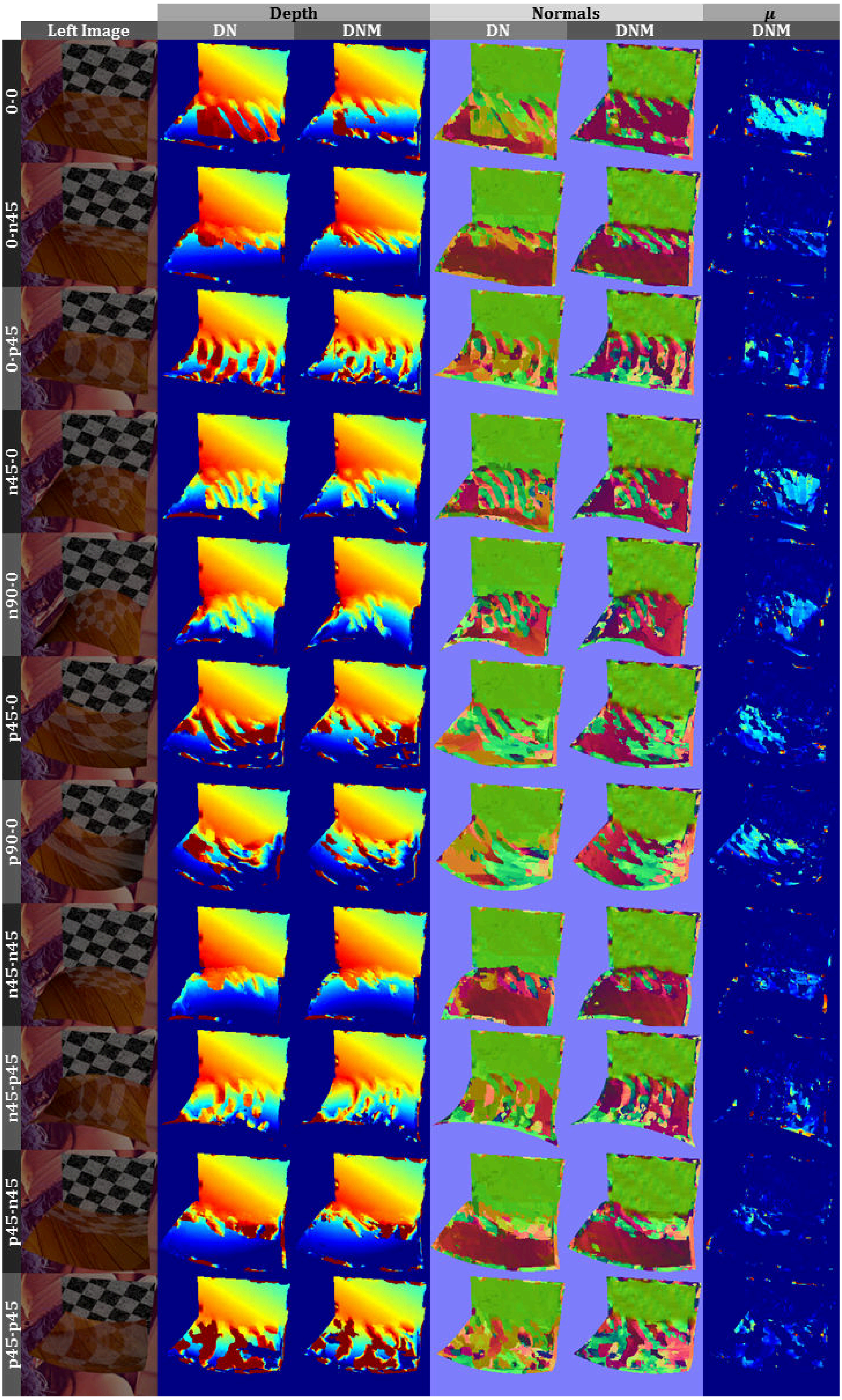


Figure 6.19: All Results for DNM, $\mu = 0.4$. Color coding as in Figure 6.16. With increasing μ, the results start to deteriorate. For some cases, the improvements can still be determined visually. The DN result for 0-0 indicates one of the issues: with the even larger choices of μ, the whole lower plane is initialized with the wrong depth, rendering the propagation of parameters from direct neighbors fruitless.

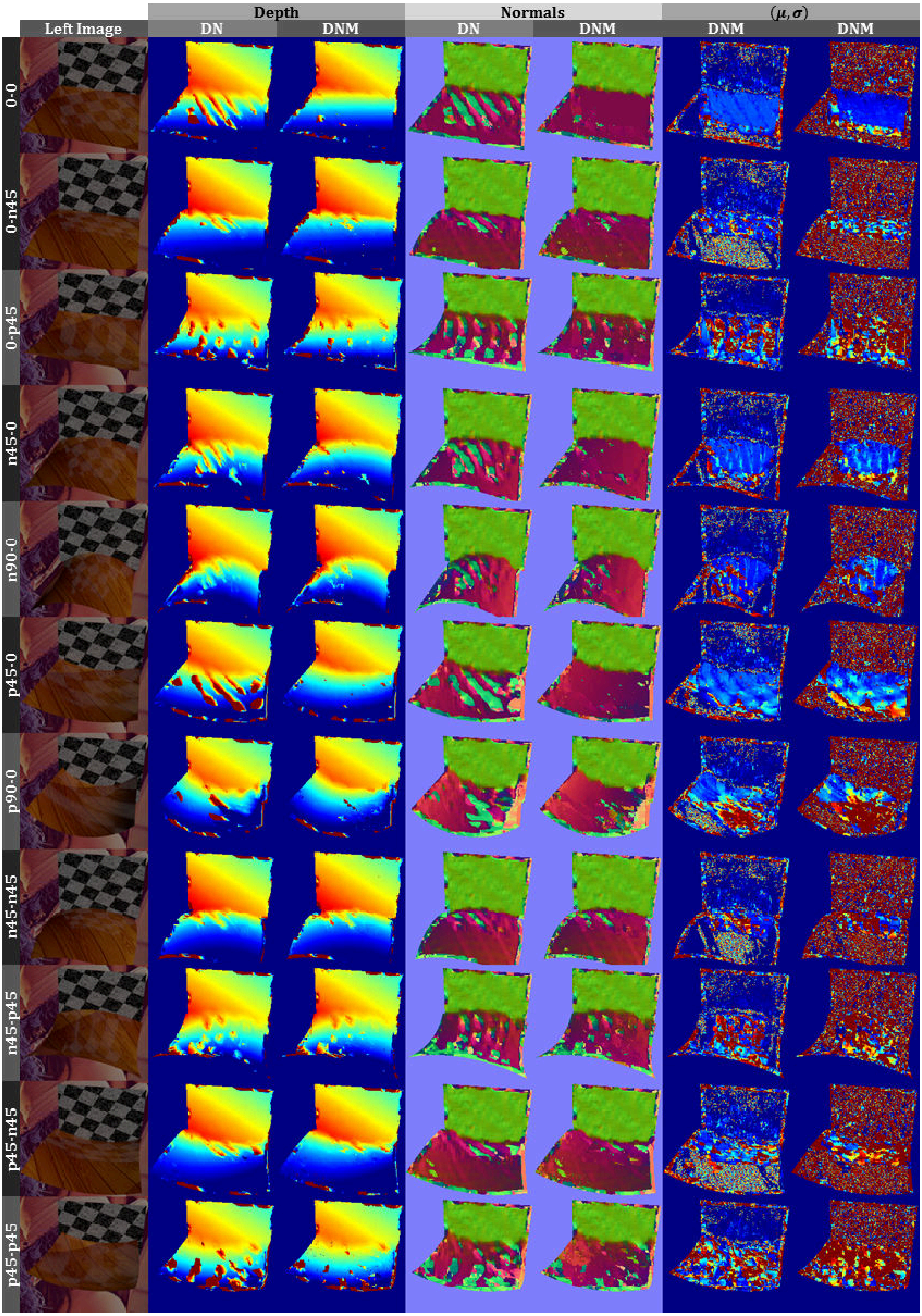


Figure 6.20: All Results for DNMS, $\mu = 0.25$, $\sigma = 0.01$. Color coding as in Figure 6.16. σ encoded in 0-0.1. Most real world specular objects are not a superposition of perfect mirror and diffuse surface, but display rough specular reflections. These can be tackled using the DNMS model, which additionally estimates the width of the specular lobe. The proposed algorithm uses a small angle approximation to compute the specular reflection such that it works best for small σ. Fortunately, for larger σ the effect of the specular reflection is diminished.

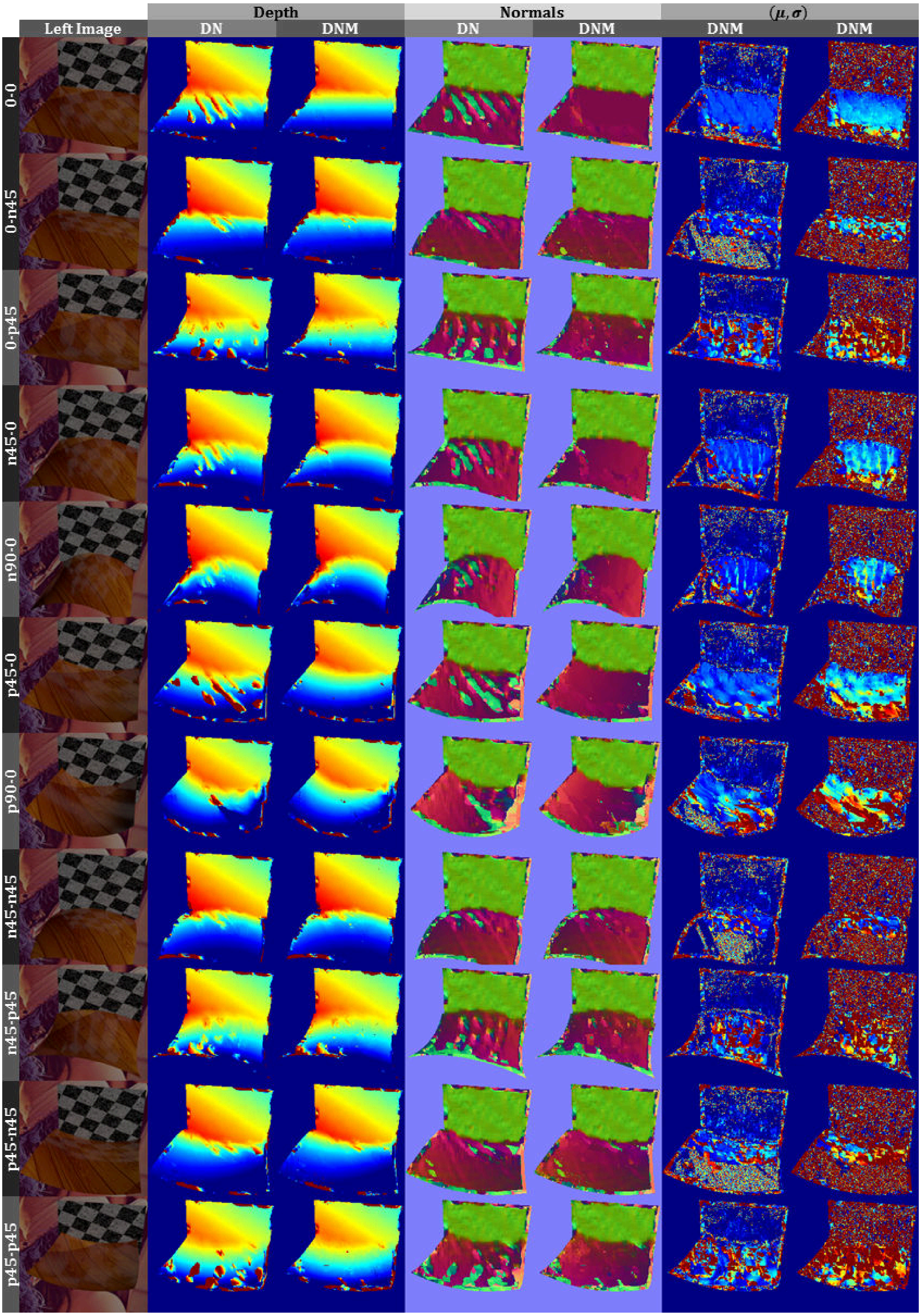


Figure 6.21: All Results for DNMS, $\mu = 0.25$, $\sigma = 0.02$. Color coding as in Figure 6.20.

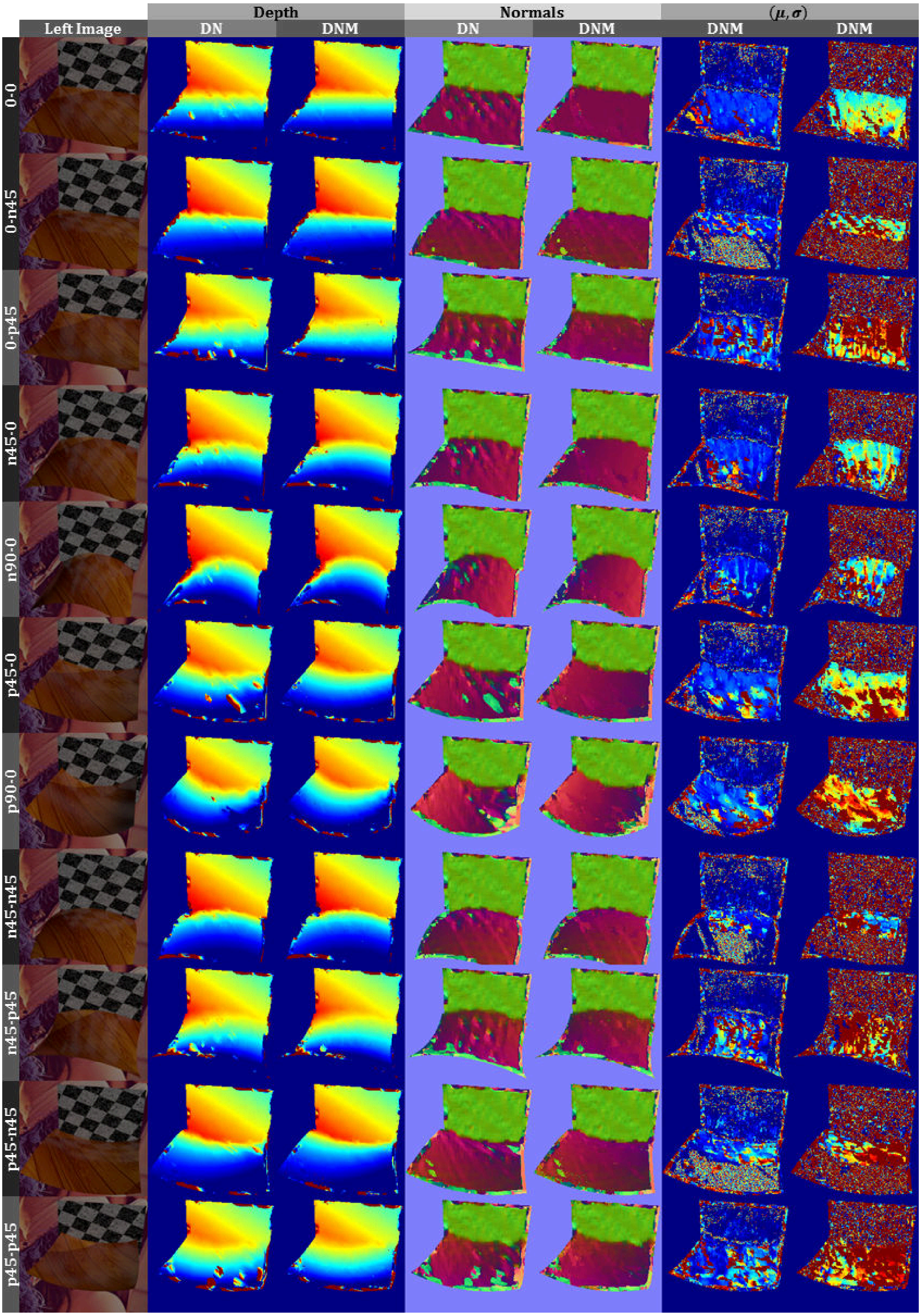


Figure 6.22: All Results for DNMS, $\mu = 0.25$, $\sigma = 0.04$. Color coding as in Figure 6.20.

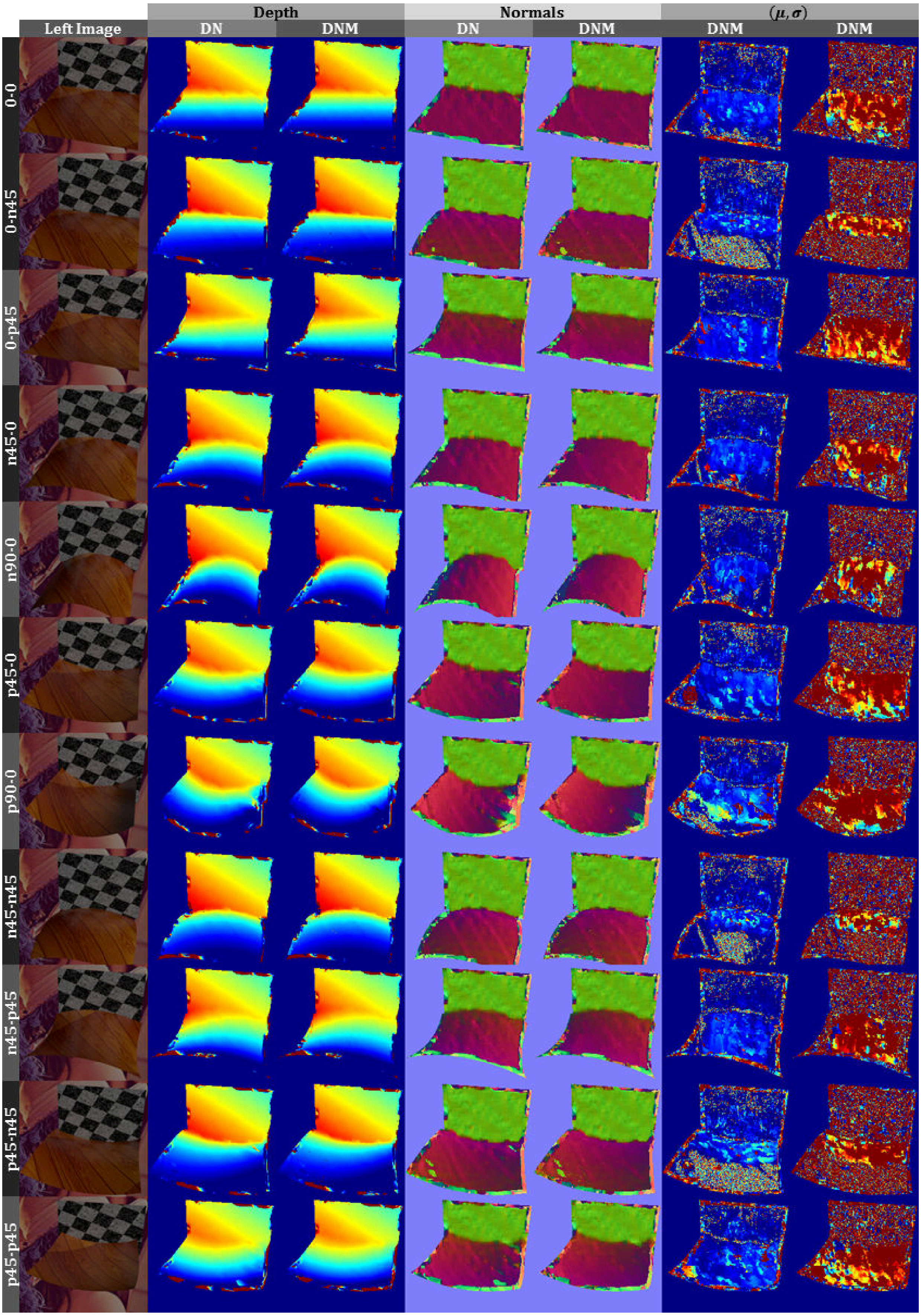


Figure 6.23: All Results for DNMS, $\mu = 0.25$, $\sigma = 0.1$. Color coding as in Figure 6.20. With even larger values of σ, the results and the appearance slowly converges to the diffuse world situation. With the DNMS model,it is still possible to estimate materials and marginally improve on the normal orientations.

6.7.2 Outlook

Towards real world imagery The first question that naturally occurs is: What about real images? Preliminary experiments suggest that there are two additional aspects that need to be taken into consideration the dynamic range of cameras and the non-linearity of the image sensor. Both effects lead to a violation of the assumption that the signal received is a linear combination of diffuse and reflected signal. A real working system therefore has to therefore operate with cameras with sufficient dynamic range and more importantly, we have to not only model light transport and camera geometry but also the sensor response.

Model Improvement The current model only explains single bounces, but reality often displays multi-reflections. How do we handle these? Will an iterative Radiosity-like [65] approach suffice, where the observed image is iteratively replaced by the current diffuse image? Or do we need to explicitly model higher order bounces. How do we handle reflections of occlusion boundaries? These areas in the image display similar edge fattening effects like normal occlusion boundaries do. Finally, how can we include (careful) regularization techniques such as PM-Huber [76] or PMBP [16] to realize real world results that improve on the local data term?

Utilizing the Scale-Space Insight The experiments with a scale-space approach gave some insights into why such techniques are not suitable for tackling reflections. The cues delivered at different scales often hint different geometries as reflections and object texture can appear at different scales. Conversely, strongly deviating results of diffuse stereo (e.g. DN-CPM), which are run independently on different scales of an observed image. could hint the presence of specular reflections and could therefore further reduce the viable search space.

Application to ToF imaging As we have seen in Chapter 3 ToF imaging also has issues in presence of specular surfaces. Therefore, it is a viable question whether the methods presented here can be extended to account for multi-path effects in ToF imaging. The standard rendering equation used in the derivation for stereo is a steady state equation and disregards the finite speed of light. An equation that takes the finite speed of light into account is called the transient render equation [168] and renderers that try to handle this are called transient renderers. Fortunately, a standard global illumination renderer can be converted into a transient renderer by making few modifications to existing algorithms [125], which suggests that algorithms that try to infer the inverse problem may have a similar overhead.

Learning Reflection Correction Trained humans can easily recognize areas in depth maps that are erroneous due to reflections. In experience, this is true for both stereo and Time-of-Flight imaging. Therefore, it seems to be a valid question for future research to investigate whether the mapping between erroneous depth and correct depth can be learned. This approach is a major departure from the 'derivation from first principles' approach presented here. Yet, to speed up the reconstruction process, this by all means could be a viable method to obtain an initial guess that is close to the real result. Note, that the approach envisaged here is somewhat different to the example-based approach mentioned earlier in [178] where the stereo matching cost is learned and not the correction of the depth map.

7 Conclusion

WITH THE BODY OF WORK PRESENTED in the previous chapters, it remains for me to conclude this thesis with a review of what was done and an outlook on what I believe is yet to come.

7.1 Summary

The starting point for the work presented was the construction of a ToF-stereo fusion system. The goal was to harness the strengths of the individual subsystems while at the same time being robust towards their respective flaws. In Chapter 3, I presented such a system and described fusion techniques [131] that make use of heuristic confidence measures derived from the input image. I also showed that these techniques display many of the desired properties on qualitative and quantitative datasets. These were a) robustness towards textureless surfaces, b) robustness towards ToF noise if scene texture is present, c) no errors due to occlusions, and d) speed of execution, while e) retaining the high resolution of the stereo camera. Yet, there remained issues in the resulting system. As so often, these issues were in fact the driving force behind the work subsequently undertaken. To summarize it, the main issues were:

- Reflective surfaces causing errors in both ToF and stereo resulting in erroneous fusion results.

- Errors due to ToF range ambiguity that (current) early fusion techniques cannot handle.
- Alignment errors due to inaccurate extrinsic calibration.
- The requirement of many hand-tuned parameters.

The first aspect was the requirement of many hand-tuned 'magic' parameters. As an engineered system and in terms of the results obtained [131, 171], this was quite acceptable. Yet, from a scientific perspective, it remained a bit unsatisfactory.

An early idea I had was to automatically estimate the parameters using learning techniques [70, 53]. While this approach has the merit of losing the heuristics, it still does not yield further insights into the reasons for specific values. Instead, I realized that the key lay in digging deeper into the matter and understanding the underlying processes. The following months of my thesis were therefore characterized by studying other fusion systems and evaluation methods [133, 132] as well as understanding the ToF measurement process [111, 106] and systematic effects that occur therein [125, 66].

In turn, this led to the establishment of the full symmetric fusion model presented in the beginning of Chapter 3. Here I showed how the majority of existing techniques derive from this model by a series of approximations. While inference of this model is still subject to future work, I believe that it is the key to improvement in future ToF-stereo fusion systems.

Further research into the least squares problem underlying the estimated ToF parameters led to the work presented in Chapter 5. This Chapter presented a method to extend the effective range of phase-based ToF cameras by changing the modulation frequency between sub-frames. The depth is then estimated by subsequently solving a modified version of the original least squares parameter estimation problem. The advantage of this method is that it can be implemented without (great) modifications to the camera hardware and that it relies on the same number of sub-frame measurements as the standard camera acquisition. Also, unlike related work, the method does not rely on strong prior assumptions on the scene composition. As a proof of concept, I displayed results on real and synthetic data that verify the claims made.

The alignment issues were revisited in Chapter 4 in a slightly different setting. Here, I presented a pipeline for the creation of large amounts of reference data for evaluating stereo matching from LIDAR measurements. The main contribution here was a rigorous analysis of how measurement errors and uncertainties in relative pose estimation between LIDAR and stereo frame introduce errors in the stereo reference data. As a result, I was able to present stereo reference data with per pixel uncertainty estimates, which can subsequently be used to improve the performance analysis of passive stereo. Another insight gained was that not all parts of the reference data are suitable for benchmarking stereo systems claiming to be sub-pixel accurate. The reason being that frequently (depth edges, areas very close to the cameras, bushes etc.) the uncertainty in the reference data is well beyond one pixel. The main areas of future work lie in model refinement, the incorporation of uncertainty analysis to other existing datasets and modalities as well as in an actual evaluation of stereo methods.

The final chapter presented (Chapter 6) was motivated by errors caused by specular surfaces in vision systems. There, I investigated how reflections affect stereo matching and also presented methods towards solving them. To this end, I revisited stereo matching as a least squares problem and derived a more general model based on the combination of the render equation and a pinhole camera model. I then showed how standard diffuse world stereo is a special case and thus derived two new models that take the first light bounce into account. Subsequently, I showed how these models can be optimized using continuous data driven PatchMatch. Results on synthetic datasets gave evidence that by modeling surface specularity it is not only possible to resolve the errors in stereo, but it is also possible to obtain material information from these surfaces. Additionally, results showed that reflections can also lead to a more accurate reconstruction of geometry due to the strong cues on surface normals that they evidence.

7.2 Outlook

A detailed outlook on each topic was given at the end of the respective chapter. Here, I will therefore take the opportunity of discussing more general aspects that I believe are of importance.

The recurring theme in my work is probably best characterized by the word ‘revisiting’. In Chapter 3 and 5, I revisited the Time-of-Flight measurement problem to gain new insights on fusion techniques and Time-of-Flight imaging itself. In Chapter 4, I revisited error propagation to gain insights on combined measurement systems and on performance analysis. Finally, in Chapter 6, I revisited camera models and light transport laws to derive novel stereo techniques.

Very often it turned out that the existing baseline models themselves are approximations of more complex ones. These approximations were often made for reasons of tractability at the time the model was first envisioned. Therefore, with the computational power and novel inference techniques available today, it may be worthwhile to again dig even deeper.

First thing to note to this end is that very few depth imaging systems make use of all available depth cues. While the exploitation of each of these cues individually (depth of field, shading, stereo, modulation, structured light) has been well researched, few systems try to jointly harness all available cues. And while each individual problem may be challenging alone, it might very well be that additional insights can be gained by looking at such joint models. It could also be that some problems that are currently being solved using sophisticated regularization techniques, naturally resolve when considering joint models.

For joint estimation in a non-heuristic manner, it is equally important to assess measurement and parameter uncertainties. These uncertainties then allow methods to decide (implicitly or explicitly) which cues to rely on in which situation. Ultimately, such systems also possess power of introspection, i.e. the ability to assess whether it is failing or not and whether the improvements made are significant at all.

Finally, an aspect that I did not touch upon in my thesis is the role of sensors in real cameras. These are typically subject to nonlinearities that may cause vision systems to fail if not accounted for. As an example, the formulas derived for range extension had to be extended to incorporate an approximation of the internal depth calibration in order to make it work on real data. Similarly, the non-linear photo response is the key for tackling reflections in real imagery.

Summarizing, I believe that there is yet a lot to be learned from holistic models that incorporate all aspects of depth imaging systems: light transport, camera properties, sensor characterization and measurement uncertainties.

List of Own Publications

[15] Berger K., Meister S., **Nair R.**, and Kondermann D. "A State of the Art Report on Kinect Sensor Setups in Computer Vision". English. In: *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications.* Ed. by Grzegorzek M., Theobalt C., Koch R., and Kolb A. Vol. 8200. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 257-272. ISBN: 978-3-642-44963-5. DOI: 10.1007/978-3-642-44964-2_12

[53] Fiaschi L., **Nair R.**, Koethe U., and Hamprecht F. "Learning to count with regression forest and structured labels". In: *Pattern Recognition (ICPR), 2012 21st International Conference on.* Nov. 2012, pp. 2685-2688. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6460719

[66] Gottfried J.-M., **Nair R.**, Meister S., Garbe C., and Kondermann D. "Time of Flight Motion Compensation Revisited". In: *Image Processing, 2014. Proceedings., International Conference on.* IEEE. 2014. URL: http://ipm.iwr.uni-heidelberg.de/publications/PDFs/2014/gottfried_ICIP2014.pdf

[70] Haeusler R., **Nair R.**, and Kondermann D. "Ensemble Learning for Confidence Measures in Stereo Vision". In: *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on.* 2013, pp. 305-312. DOI: 10.1109/CVPR.2013.46

[102] Kondermann D., **Nair R.**, Meister S., Mischler W., Güssefeld B., Honauer K., Sabine H., Brenner C., and Jähne B. "Stereo Ground Truth With Error Bars". In: *Proc. ACCV.* Nov. 2014. URL: http://hci.iwr.uni-heidelberg.de/Benchmarks/document/StereoErrorBars

[106] Lefloch D., **Nair R.**, Lenzen F., Schäfer H., Streeter L., Cree M. J., Koch R., and Kolb A. "Technical Foundation and Calibration Methods for Time-of-Flight Cameras". English. In: *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications.* Ed. by Grzegorzek M., Theobalt C., Koch R., and Kolb A. Vol. 8200. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 3-24. ISBN: 978-3-642-44963-5. DOI: 10.1007/978-3-642-44964-2_1

[111] Lenzen F., Kim K. I., Schäfer H., **Nair R.**, Meister S., Becker F., Garbe C. S., and Theobalt C. "Denoising Strategies for Time-of-Flight Data". English. In: *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications.* Ed. by Grzegorzek M., Theobalt C., Koch R., and Kolb A. Vol. 8200. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 25-45. ISBN: 978-3-642-44963-5. DOI: 10.1007/978-3-642-44964-2_2

[125] Meister S., **Nair R.**, and Kondermann D. "Simulation of Time-of-Flight Sensors using Global Illumination". In: *Vision, Modeling and Visualization.* Ed. by Bronstein M., Favre J., and Hormann K. The Eurographics Association, 2013. ISBN: 978-3-905674-51-4. DOI: 10.2312/PE.VMV.VMV13.033-040

[131] **Nair R.**, Lenzen F., Meister S., Schäfer H., Garbe C., and Kondermann D. "High Accuracy TOF and Stereo Sensor Fusion at Interactive Rates". English. In: *ECCV 2012. Workshops and Demonstrations.* Ed. by Fusiello A., Murino V., and Cucchiara R. Vol. 7584. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, pp. 1-11. ISBN: 978-3-642-33867-0. DOI: 10.1007/978-3-642-33868-7_1

[133] **Nair R.**, Ruhl K., Lenzen F., Meister S., Schäfer H., Garbe C. S., Eisemann M., Magnor M., and Kondermann D. "A Survey on Time-of-Flight Stereo Fusion". English. In: *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications.* Ed. by Grzegorzek M., Theobalt C., Koch R., and Kolb A. Vol. 8200. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 105-127. ISBN: 978-3-642-44963-5. DOI: 10.1007/978-3-642-44964-2_6

[132] **Nair R.**, Meister S., Lambers M., Balda M., Hofmann H., Kolb A., Kondermann D., and Jähne B. "Ground Truth for Evaluating Time of Flight Imaging". English. In: *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications.* Ed. by Grzegorzek M., Theobalt C., Koch R., and Kolb A. Vol. 8200. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 52-74. ISBN: 978-3-642-44963-5. DOI: 10.1007/978-3-642-44964-2_4

[171] Spielmann S., Helzle V., and **Nair R.** "On-set Depth Capturing for VFX Productions Using Time of Flight". In: *ACM SIGGRAPH 2013 Talks.* SIGGRAPH '13. Anaheim, California: ACM, 2013, 13:1-13:1. ISBN: 978-1-4503-2344-4. DOI: 10.1145/2504459.2504475

List of Tables

List of Figures

Bibliography

[1] Abraham S. and Hau T. "Towards autonomous high-precision calibration of digital cameras". In: vol. 3174. 1997, pp. 82–93. DOI: 10.1117/12.279802.

[2] Adato Y., Vasilyev Y., Zickler T., and Ben-Shahar O. "Shape from Specular Flow". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 32.11 (Nov. 2010), pp. 2054–2070. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2010.126.

[3] Afonso M., Bioucas-Dias J., and Figueiredo M. "Fast Image Recovery Using Variable Splitting and Constrained Optimization". In: *Image Processing, IEEE Transactions on* 19.9 (Sept. 2010), pp. 2345–2356. ISSN: 1057-7149. DOI: 10.1109/TIP.2010.2047910.

[4] Agarwal S., Mierle K., and Others. *Ceres Solver*. http://ceres-solver.org. 2014.

[5] Alvarez L., Sánchez J., and Weickert J. "A Scale-Space Approach to Nonlocal Optical Flow Calculations". In: *Scale-Space Theories in Computer Vision*. Ed. by Nielsen M., Johansen P., Olsen O., and Weickert J. Vol. 1682. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 1999, pp. 235–246. ISBN: 978-3-540-66498-7. DOI: 10.1007/3-540-48236-9_21.

[6] Ayers G. R. and Dainty J. C. "Iterative blind deconvolution method and its applications". In: *Opt. Lett.* 13.7 (July 1988), pp. 547–549. DOI: 10.1364/OL.13.000547.

[7] Badino H., Franke U., and Pfeiffer D. "The Stixel World - A Compact Medium Level Representation of the 3D-World". In: *Pattern Recognition*. Ed. by Denzler J., Notni G., and Herbert S. Vol. 5748. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2009, pp. 51–60. ISBN: 978-3-642-03797-9. DOI: 10.1007/978-3-642-03798-6_6.

[8] Baillard C. and Maıtre H. "3-D Reconstruction of Urban Scenes from Aerial Stereo Imagery: A Focusing Strategy". In: *Computer Vision and Image Understanding* 76.3 (1999), pp. 244 –258. ISSN: 1077-3142. DOI: http://dx.doi.org/10.1006/cviu.1999.0793.

[9] Bajcsy R., Lee S., and Leonardis A. "Color image segmentation with detection of highlights and local illumination induced by inter-reflections". In: *Pattern Recognition, 1990. Proceedings., 10th International Conference on*. Vol. i. June 1990, pp. 785–790. DOI: 10.1109/ICPR.1990.118217.

[10] Baker S., Scharstein D., Lewis J., Roth S., Black M. J., and Szeliski R. "A Database and Evaluation Methodology for Optical Flow". In: *International Journal of Computer Vision* 92.1 (2011), pp. 1–31. ISSN: 0920-5691. DOI: 10.1007/s11263-010-0390-2.

[11] Barnes C., Shechtman E., Finkelstein A., and Goldman D. B. "PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing". In: *ACM Trans. Graph.* 28.3 (July 2009), 24:1–24:11. ISSN: 0730-0301. DOI: 10.1145/1531326.1531330.

[12] Barron J. and Malik J. “Shape, albedo, and illumination from a single image of an unknown object”. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.* June 2012, pp. 334–341. DOI: 10.1109/CVPR.2012.6247693.

[13] Bartczak B. and Koch R. “Dense Depth Maps from Low Resolution Time-of-Flight Depth and High Resolution Color Views”. In: Lecture Notes in Computer Science 5876 (2009). Ed. by Bebis G., Boyle R., Parvin B., Koracin D., Kuno Y., Wang J., Pajarola R., Lindstrom P., Hinkenjann A., Encarnacao M. L., Silva C. T., and Coming D., pp. 228–239. DOI: 10.1007/978-3-642-10520-3_21.

[14] Beder C., Bartczak B., and Koch R. “A Combined Approach for Estimating Patchlets from PMD Depth Images and Stereo Intensity Images”. In: *Pattern Recognition.* Ed. by Hamprecht F. A., Schnörr C., and Jähne B. Vol. 4713. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2007, pp. 11–20. ISBN: 978-3-540-74933-2. DOI: 10.1007/978-3-540-74936-3_2.

[15] Berger K., Meister S., Nair R., and Kondermann D. “A State of the Art Report on Kinect Sensor Setups in Computer Vision”. In: *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications.* Ed. by Grzegorzek M., Theobalt C., Koch R., and Kolb A. Vol. 8200. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 257–272. ISBN: 978-3-642-44963-5. DOI: 10.1007/978-3-642-44964-2_12.

[16] Besse F., Rother C., Fitzgibbon A., and Kautz J. “PMBP: PatchMatch Belief Propagation for Correspondence Field Estimation”. In: *International Journal of Computer Vision* 110.1 (2014), pp. 2–13. ISSN: 0920-5691. DOI: 10.1007/s11263-013-0653-9.

[17] Blake A. and Brelstaff G. “Geometry From Specularities”. In: (Dec. 1988), pp. 394–403. DOI: 10.1109/CCV.1988.590016.

[18] Blender Online Community. *Blender - a 3D modelling and rendering package.* Blender Institute, Amsterdam: Blender Foundation, 2014. URL: http://www.blender.org.

[19] Bleyer M., Rhemann C., and Rother C. “PatchMatch Stereo - Stereo Matching with Slanted Support Windows”. In: *Proceedings of the British Machine Vision Conference.* BMVA Press, 2011, pp. 14.1–14.11. ISBN: 1-901725-43-X. DOI: 10.5244/C.25.14.

[20] Boehler W., Bordas-Vicent M, and Marbs A. “Investigating laser scanner accuracy”. In: *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 34.Part 5 (2003), pp. 696–701. URL: http://i3mainz.de/sites/default/files/public/data/laserscanner_accuracy.pdf.

[21] Bolles R. C., Baker H. H., and Marimont D. H. “Epipolar-plane image analysis: An approach to determining structure from motion”. In: *International Journal of Computer Vision* 1.1 (1987), pp. 7–55. ISSN: 0920-5691. DOI: 10.1007/BF00128525.

[22] Bosch A., Zisserman A., and Muoz X. “Image Classification using Random Forests and Ferns”. In: *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on.* Oct. 2007, pp. 1–8. DOI: 10.1109/ICCV.2007.4409066.

[23] Boyd S. and Vandenberghe L. *Convex Optimization.* New York, NY, USA: Cambridge University Press, 2004. ISBN: 0521833787. URL: http://stanford.edu/~boyd/cvxbook/.

[24] Boykov Y., Veksler O., and Zabih R. “Fast approximate energy minimization via graph cuts”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 23.11 (Nov. 2001), pp. 1222–1239. ISSN: 0162-8828. DOI: 10.1109/34.969114.

[25] Bradski G. “The OpenCV Library”. In: *Dr. Dobb’s Journal of Software Tools* (2000). URL: http://opencv.org.

[26] Breiman L. “Random Forests”. In: *Machine Learning* 45.1 (2001), pp. 5–32. ISSN: 0885-6125. DOI: 10.1023/A:1010933404324.

[27] Brelstaff G. and Blake A. “Detecting Specular Reflections Using Lambertian Constraints”. In: *Computer Vision., Second International Conference on.* Dec. 1988, pp. 297–302. DOI: 10.1109/CCV.1988.590004.

[28] Bresenham J. E. “Algorithm for Computer Control of a Digital Plotter”. In: *IBM Syst. J.* 4.1 (Mar. 1965), pp. 25–30. ISSN: 0018-8670. DOI: 10.1147/sj.41.0025. URL: http://dx.doi.org/10.1147/sj.41.0025.

[29] Buades A., Coll B., and Morel J. “A Review of Image Denoising Algorithms, with a New One”. In: *Multiscale Modeling & Simulation* 4.2 (2005), pp. 490–530. DOI: 10.1137/040616024. eprint: http://dx.doi.org/10.1137/040616024.

[30] Butler D. J., Wulff J., Stanley G. B., and Black M. J. “A Naturalistic Open Source Movie for Optical Flow Evaluation”. In: *Computer Vision, ECCV 2012.* Ed. by Fitzgibbon A., Lazebnik S., Perona P., Sato Y., and Schmid C. Vol. 7577. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, pp. 611–625. ISBN: 978-3-642-33782-6. DOI: 10.1007/978-3-642-33783-3_44.

[31] Castaneda V., Mateus D., and Navab N. “Stereo time-of-flight”. In: *Computer Vision (ICCV), 2011 IEEE International Conference on.* Nov. 2011, pp. 1684–1691. DOI: 10.1109/ICCV.2011.6126431.

[32] Cech J. and Sara R. “Efficient Sampling of Disparity Space for Fast And Accurate Matching”. In: *Computer Vision and Pattern Recognition, 2007. CVPR ’07. IEEE Conference on.* June 2007, pp. 1–8. DOI: 10.1109/CVPR.2007.383355.

[33] Chambolle A. and Pock T. “A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging”. In: *Journal of Mathematical Imaging and Vision* 40.1 (2011), pp. 120–145. ISSN: 0924-9907. DOI: 10.1007/s10851-010-0251-1.

[34] Chetverikov D., Svirko D., Stepanov D., and Krsek P. “The Trimmed Iterative Closest Point algorithm”. In: *Pattern Recognition, 2002. Proceedings. 16th International Conference on.* Vol. 3. 2002, pp. 545–548. DOI: 10.1109/ICPR.2002.1047997.

[35] Choi O. and Lee S. “Wide range stereo time-of-flight camera”. In: *Image Processing (ICIP) 2012, 19th IEEE International Conference on.* Sept. 2012, pp. 557–560. DOI: 10.1109/ICIP.2012.6466920.

[36] Choi O., Lim H., Kang B., Kim Y. S., Lee K., Kim J., and Kim C.-Y. “Range unfolding for Time-of-Flight depth cameras”. In: *Image Processing (ICIP) 2010, 17th IEEE International Conference on.* Sept. 2010, pp. 4189–4192. DOI: 10.1109/ICIP.2010.5651383.

[37] Clason C., Jin B., and Kunisch K. “A Duality-Based Splitting Method for L1-TV Image Restoration with Automatic Regularization Parameter Choice”. In: *SIAM Journal on Scientific Computing* 32.3 (2010), pp. 1484–1505. DOI: 10.1137/090768217.

[38] Cook R. L. and Torrance K. E. “A Reflectance Model for Computer Graphics”. In: *ACM Trans. Graph.* 1.1 (Jan. 1982), pp. 7–24. ISSN: 0730-0301. DOI: 10.1145/357290.357293.

[39] Criminisi A., Kang S. B., Swaminathan R., Szeliski R., and Anandan P. “Extracting layers and analyzing their specular properties using epipolar-plane-image analysis”. In: *Computer Vision and Image Understanding* 97.1 (2005), pp. 51 –85. ISSN: 1077-3142. DOI: 10.1016/j.cviu.2004.06.001.

[40] Crow F. C. “Summed-area Tables for Texture Mapping”. In: *SIGGRAPH Comput. Graph.* 18.3 (Jan. 1984), pp. 207–212. ISSN: 0097-8930. DOI: 10.1145/964965.808600.

[41] Dal Mutto C., Zanuttigh P., and Cortelazzo G. M. “A Probabilistic Approach to ToF and Stereo Data Fusion”. In: *3DPVT.* Paris, France, May 2010. URL: http://lttm.dei.unipd.it/nuovo/Papers/10_3DPVT.pdf.

[42] Dal Mutto C., Zanuttigh P., Mattoccia S., and Cortelazzo G. “Locally Consistent ToF and Stereo Data Fusion”. In: *Computer Vision, ECCV 2012. Workshops and Demonstrations.* Ed. by Fusiello A., Murino V., and Cucchiara R. Vol. 7583. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, pp. 598–607. ISBN: 978-3-642-33862-5. DOI: 10.1007/978-3-642-33863-2_62.

[43] Davidson D. B. *Computational electromagnetics for RF and microwave engineering.* Cambridge University Press, 2005. ISBN: 978-0-521-51891-8. URL: http://www.cambridge.org/us/academic/subjects/engineering/rf-and-microwave-engineering/computational-electromagnetics-rf-and-microwave-engineering-2nd-edition.

[44] Davison A., Reid I., Molton N., and Stasse O. “MonoSLAM: Real-Time Single Camera SLAM”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 29.6 (June 2007), pp. 1052–1067. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2007.1049.

[45] Dickscheid T., Läbe T., and Förstner W. “Benchmarking automatic bundle adjustment results”. In: *XXI. ISPRS congress, Beijing.* ISPRS. 2008. URL: http://www.isprs.org/proceedings/XXXVII/congress/3_pdf/02.pdf.

[46] Diego F., Ponsa D., Serrat J., and Lopez A. “Video Alignment for Change Detection”. In: *Image Processing, IEEE Transactions on* 20.7 (July 2011), pp. 1858–1869. ISSN: 1057-7149. DOI: 10.1109/TIP.2010.2095873.

[47] Donath A. and Kondermann D. “Is Crowdsourcing for Optical Flow Ground Truth Generation Feasible?” In: Lecture Notes in Computer Science 7963 (2013). Ed. by Chen M., Leibe B., and Neumann B., pp. 193–202. DOI: 10.1007/978-3-642-39402-7_20.

[48] EMVA 1288 Working Group. *EMVA Standard 1288 - Standard for Characterization of Image Sensors and Cameras*. 2010. DOI: 10.5281/zenodo.10696.

[49] Erz M. and Jähne B. "Radiometric and spectrometric calibrations, and distance noise measurement of TOF cameras". In: *3rd Workshop on Dynamic 3-D Imaging*. Ed. by Koch R. and Kolb A. Vol. 5742. Lecture Notes in Computer Science. Springer, 2009, pp. 28–41. DOI: 10.1007/978-3-642-03778-8_3.

[50] Everingham M., Van Gool L., Williams C. K., Winn J., and Zisserman A. "The Pascal Visual Object Classes (VOC) Challenge". In: *International Journal of Computer Vision* 88.2 (2010), pp. 303–338. ISSN: 0920-5691. DOI: 10.1007/s11263-009-0275-4.

[51] Falie D. and Buzuloiu V. "Wide range Time of Flight camera for outdoor surveillance". In: *Microwaves, Radar and Remote Sensing Symposium, 2008*. Sept. 2008, pp. 79–82. DOI: 10.1109/MRRS.2008.4669550.

[52] Felzenszwalb P. F. and Huttenlocher D. P. "Efficient Graph-Based Image Segmentation". In: *International Journal of Computer Vision* 59.2 (2004), pp. 167–181. ISSN: 0920-5691. DOI: 10.1023/B:VISI.0000022288.19776.77.

[53] Fiaschi L., Nair R., Koethe U., and Hamprecht F. "Learning to count with regression forest and structured labels". In: *Pattern Recognition (ICPR), 2012 21st International Conference on*. Nov. 2012, pp. 2685–2688. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6460719.

[54] Fischer J., Arbeiter G., and Verl A. "Combination of Time-of-Flight depth and stereo using semiglobal optimization". In: *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. May 2011, pp. 3548–3553. DOI: 10.1109/ICRA.2011.5979999.

[55] Förstner W. "Reliability analysis of parameter estimation in linear models with applications to mensuration problems in computer vision". In: *Computer Vision, Graphics, and Image Processing* 40.3 (1987), pp. 273 –310. ISSN: 0734-189X. DOI: 10.1016/S0734-189X(87)80144-5.

[56] Freedman D., Krupka E., Smolin Y., Leichter I., and Schmidt M. "SRA: Fast Removal of General Multipath for ToF Sensors". In: *CoRR* abs/1403.5919 (2014). URL: http://arxiv.org/abs/1403.5919.

[57] Frey B. J., Koetter R., and Petrovic N. "Very loopy belief propagation for unwrapping phase images". In: *Advances in Neural Information Processing Systems 14*. Ed. by Dietterich T., Becker S., and Ghahramani Z. MIT Press, 2002, pp. 737–743. URL: http://papers.nips.cc/paper/2126-very-loopy-belief-propagation-for-unwrapping-phase-images.pdf.

[58] Fua P. "Combining stereo and monocular information to compute dense depth maps that preserve depth discontinuities". In: *In Proceedings of the 12th International Joint Conference on Artificial Intelligence*. 1991, pp. 1292–1298. URL: http://www.ijcai.org/Past%20Proceedings/IJCAI-91-VOL2/PDF/097.pdf.

[59] Gandhi V., Cech J., and Horaud R. "High-resolution depth maps based on TOF-stereo fusion". In: *Robotics and Automation (ICRA), 2012 IEEE International Conference on.* May 2012, pp. 4742–4749. DOI: 10.1109/ICRA.2012.6224771.

[60] Geiger A., Lenz P., and Urtasun R. "Are we ready for autonomous driving? The KITTI vision benchmark suite". In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.* June 2012, pp. 3354–3361. DOI: 10.1109/CVPR.2012.6248074.

[61] Gershon R., Jepson A. D., and Tsotsos J. K. "The Use of Color in Highlight Identification." In: *IJCAI.* 1987, pp. 752–754. URL: http://ijcai.org/Past%20Proceedings/IJCAI-87-VOL2/PDF/034.pdf.

[62] Girshick R., Felzenszwalb P., and McAllester D. "Discriminatively trained deformable part models, release 5". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2012). DOI: 10.1109/CVPR.2008.4587597.

[63] Goldstein R. M., Zebker H. A., and Werner C. L. "Satellite radar interferometry: Two-dimensional phase unwrapping". In: *Radio Science* 23.4 (1988), pp. 713–720. DOI: 10.1029/RS023i004p00713.

[64] Goodman J. *Introduction to Fourier optics.* McGraw-Hill, 2008. URL: http://cds.cern.ch/record/108347.

[65] Goral C. M., Torrance K. E., Greenberg D. P., and Battaile B. "Modeling the Interaction of Light Between Diffuse Surfaces". In: *SIGGRAPH Comput. Graph.* 18.3 (Jan. 1984), pp. 213–222. ISSN: 0097-8930. DOI: 10.1145/964965.808601.

[66] Gottfried J.-M., Nair R., Meister S., Garbe C., and Kondermann D. "Time of Flight Motion Compensation Revisited". In: *Image Processing, 2014. Proceedings., International Conference on.* IEEE. 2014. URL: http://ipm.iwr.uni-heidelberg.de/publications/PDFs/2014/gottfried_ICIP2014.pdf.

[67] Grzegorzek M., Theobalt C., Koch R., and Kolb A. *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications: Dagstuhl 2012 Seminar on Time-of-Flight Imaging and 2013 Workshop on Imaging New Modalities.* Springer Berlin Heidelberg, 2013. DOI: 10.1007/978-3-642-44964-2.

[68] Gudmundsson S., Aanaes H., and Larsen R. "Fusion of stereo vision and time-of-flight imaging for improved 3d estimation". In: *IJISTA* 5.3 (2008), pp. 425–433. DOI: 10.1504/IJISTA.2008.021305.

[69] Güssefeld B., Kondermann D., Schwartz C., and Klein R. "Are reflectance field renderings appropriate for optical flow evaluation?" In: *IEEE International Conference on Image Processing (ICIP).* IEEE. Paris, France, Oct. 2014.

[70] Haeusler R., Nair R., and Kondermann D. "Ensemble Learning for Confidence Measures in Stereo Vision". In: *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on.* June 2013, pp. 305–312. DOI: 10.1109/CVPR.2013.46.

[71] Hahne U. and Alexa M. "Combining time-of-flight depth and stereo images without accurate extrinsic calibration". In: *IJISTA* 5.3 (2008), pp. 325–333. DOI: 10.1504/IJISTA.2008.021295.

[72] Hahne U. and Alexa M. "Depth Imaging by Combining Time-of-Flight and On-Demand Stereo". In: Lecture Notes in Computer Science 5742 (2009). Ed. by Kolb A. and Koch R., pp. 70–83. DOI: 10.1007/978-3-642-03778-8_6.

[73] Haltakov V., Unger C., and Ilic S. "Framework for Generation of Synthetic Ground Truth Data for Driver Assistance Applications". In: *Pattern Recognition.* Ed. by Weickert J., Hein M., and Schiele B. Vol. 8142. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 323–332. ISBN: 978-3-642-40601-0. DOI: 10.1007/978-3-642-40602-7_35.

[74] Hansard M., Lee S., Choi O., and Horaud R. *Alignment of Time-of-Flight and Stereoscopic Data.* SpringerBriefs in Computer Science. Springer London, 2013, pp. 59–75. ISBN: 978-1-4471-4657-5. DOI: 10.1007/978-1-4471-4658-2_4.

[75] Hartley R. I. and Zisserman A. *Multiple View Geometry in Computer Vision.* Second Edition. Cambridge University Press, 2004. URL: http://www.robots.ox.ac.uk/~vgg/hzbook/.

[76] Heise P., Klose S., Jensen B., and Knoll A. "PM-Huber: PatchMatch with Huber Regularization for Stereo Matching". In: *Computer Vision (ICCV), 2013 IEEE International Conference on.* Dec. 2013, pp. 2360–2367. DOI: 10.1109/ICCV.2013.293.

[77] Hirschmuller H. "Stereo Processing by Semiglobal Matching and Mutual Information". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 30.2 (Feb. 2008), pp. 328–341. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2007.1166.

[78] Hirschmuller H. and Scharstein D. "Evaluation of Cost Functions for Stereo Matching". In: *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on.* June 2007, pp. 1–8. DOI: 10.1109/CVPR.2007.383248.

[79] Hirschmuller H. and Scharstein D. "Evaluation of Stereo Matching Costs on Images with Radiometric Differences". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31.9 (Sept. 2009), pp. 1582–1599. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2008.221.

[80] Hong D., Lee H., Kim M. Y., Cho H., and Moon J. I. "Sensor fusion of phase measuring profilometry and stereo vision for three-dimensional inspection of electronic components assembled on printed circuit boards". In: *Appl. Opt.* 48.21 (July 2009), pp. 4158–4169. DOI: 10.1364/AO.48.004158.

[81] Hornacek M., Besse F., Kautz J., Fitzgibbon A., and Rother C. "Highly Overparameterized Optical Flow Using PatchMatch Belief Propagation". In: *Computer Vision , ECCV 2014.* Ed. by Fleet D., Pajdla T., Schiele B., and Tuytelaars T. Vol. 8691. Lecture Notes in Computer Science. Springer International Publishing, 2014, pp. 220–234. ISBN: 978-3-319-10577-2. DOI: 10.1007/978-3-319-10578-9_15.

[82] Hubel P. M., Liu J., and Guttosch R. J. “Spatial frequency response of color image sensors: Bayer color filters and Foveon X3”. In: vol. 5301. 2004, pp. 402–407. DOI: 10.1117/12.561568.

[83] Huhle B., Fleck S., and Schilling A. “Integrating 3D Time-of-Flight Camera Data and High Resolution Images for 3DTV Applications”. In: *3DTV Conference, 2007.* May 2007, pp. 1–4. DOI: 10.1109/3DTV.2007.4379472.

[84] Jähne B. “EMVA 1288 standard for machine vision – objective specification of vital camera data”. In: *Optik & Photonik* 5 (2010), pp. 53–54. DOI: 10.1002/opph.201190082.

[85] Jähne B. *Digitale Bildverarbeitung.* Springer-Verlag, 2012. DOI: 10.1007/978-3-642-04952-1.

[86] Jancsary J., Nowozin S., Sharp T., and Rother C. “Regression Tree Fields: An efficient, non-parametric approach to image labeling problems”. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.* June 2012, pp. 2376–2383. DOI: 10.1109/CVPR.2012.6247950.

[87] Jiminez D., Pizarro D., Mazo M., and Palazuelos S. “Modeling and correction of multipath interference in time of flight cameras”. In: *Image and Vision Computing* 32.1 (2014), pp. 1 –13. ISSN: 0262-8856. DOI: 10.1016/j.imavis.2013.10.008.

[88] Jin H., Soatto S., and Yezzi A. “Multi-view stereo beyond Lambert”. In: *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on.* Vol. 1. June 2003, pp. 171–178. DOI: 10.1109/CVPR.2003.1211351.

[89] Jin H., Yezzi A. J., and Soatto S. “Variational Multiframe Stereo in the Presence of Specular Reflections”. In: Los Alamitos, CA, USA: IEEE Computer Society, 2002, p. 626. ISBN: 0-7695-1521-5. DOI: 10.1109/TDPVT.2002.1024128.

[90] Jolivet V., Plemenos D., and Poulingeas P. “Inverse Direct Lighting with a Monte Carlo Method and Declarative Modelling”. In: *Computational Science, ICCS 2002.* Ed. by Sloot P., Hoekstra A., Tan C., and Dongarra J. Vol. 2330. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2002, pp. 3–12. ISBN: 978-3-540-43593-8. DOI: 10.1007/3-540-46080-2_1.

[91] Junker A., Stenau T., and Brenner K.-H. “Scalar wave-optical reconstruction of plenoptic camera images”. In: *Appl. Opt.* 53.25 (Sept. 2014), pp. 5784–5790. DOI: 10.1364/AO.53.005784.

[92] Kajiya J. T. “The Rendering Equation”. In: *SIGGRAPH Comput. Graph.* 20.4 (Aug. 1986), pp. 143–150. ISSN: 0097-8930. DOI: 10.1145/15886.15902.

[93] Kanade T. and Okutomi M. “A stereo matching algorithm with an adaptive window: theory and experiment”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 16.9 (Sept. 1994), pp. 920–932. ISSN: 0162-8828. DOI: 10.1109/34.310690.

[94] Kanade T., Yoshida A., Oda K., Kano H., and Tanaka M. “A stereo machine for video-rate dense depth mapping and its new applications”. In: *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR ’96, 1996 IEEE Computer Society Conference on.* June 1996, pp. 196–202. DOI: 10.1109/CVPR.1996.517074.

[95] Kanatani K. “Uncertainty modeling and model selection for geometric inference”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 26.10 (Oct. 2004), pp. 1307–1319. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2004.93.

[96] Kanatani K. “Statistical Optimization for Geometric Fitting: Theoretical Accuracy Bound and High Order Error Analysis”. In: *International Journal of Computer Vision* 80.2 (2008), pp. 167–188. ISSN: 0920-5691. DOI: 10.1007/s11263-007-0098-0.

[97] Kappes J., Andres B., Hamprecht F., Schnorr C., Nowozin S., Batra D., Kim S., Kausler B., Lellmann J., Komodakis N., and Rother C. “A Comparative Study of Modern Inference Techniques for Discrete Energy Minimization Problems”. In: *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on.* June 2013, pp. 1328–1335. DOI: 10.1109/CVPR.2013.175.

[98] Kim Y. M., Theobalt C., Diebel J., Kosecka J., Miscusik B., and Thrun S. “Multi-view image and ToF sensor fusion for dense 3D reconstruction”. In: *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on.* Sept. 2009, pp. 1542–1549. DOI: 10.1109/ICCVW.2009.5457430.

[99] Kimmel R. “Demosaicing: image reconstruction from color CCD samples”. In: *Image Processing, IEEE Transactions on* 8.9 (Sept. 1999), pp. 1221–1228. ISSN: 1057-7149. DOI: 10.1109/83.784434.

[100] Koenderink J. J. and Doorn A. J. van. “Affine structure from motion”. In: *J. Opt. Soc. Am. A* 8.2 (Feb. 1991), pp. 377–385. DOI: 10.1364/JOSAA.8.000377.

[101] Kondermann D. “Ground Truth Design Principles: An Overview”. In: *Proceedings of the International Workshop on Video and Image Ground Truth in Computer Vision Applications.* VIGTA ’13. St. Petersburg, Russia: ACM, 2013, 5:1–5:4. ISBN: 978-1-4503-2169-3. DOI: 10.1145/2501105.2501114.

[102] Kondermann D., Nair R., Meister S., Mischler W., Güssefeld B., Honauer K., Sabine H., Brenner C., and Jähne B. “Stereo Ground Truth With Error Bars”. In: *Proc. ACCV.* Nov. 2014. URL: http://hci.iwr.uni-heidelberg.de/Benchmarks/document/StereoErrorBars.

[103] Köthe U. *Generic programming for computer vision: The vigra computer vision library.* 2011. URL: https://ukoethe.github.io/vigra/.

[104] Kuhnert K. and Stommel M. “Fusion of Stereo-Camera and PMD-Camera Data for Real-Time Suited Precise 3D Environment Reconstruction”. In: *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on.* Oct. 2006, pp. 4780–4785. DOI: 10.1109/IROS.2006.282349.

[105] Lee S. W. and Bajcsy R. “Detection of specularity using color and multiple views”. In: *Computer Vision ECCV’92*. Ed. by Sandini G. Vol. 588. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 1992, pp. 99–114. ISBN: 978-3-540-55426-4. DOI: 10.1007/3-540-55426-2_13.

[106] Lefloch D., Nair R., Lenzen F., Schäfer H., Streeter L., Cree M. J., Koch R., and Kolb A. “Technical Foundation and Calibration Methods for Time-of-Flight Cameras”. In: *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*. Ed. by Grzegorzek M., Theobalt C., Koch R., and Kolb A. Vol. 8200. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 3–24. ISBN: 978-3-642-44963-5. DOI: 10.1007/978-3-642-44964-2_1.

[107] Lellmann J., Balzer J., Rieder A., and Beyerer J. “Shape from Specular Reflection and Optical Flow”. In: *International Journal of Computer Vision* 80.2 (2008), pp. 226–241. ISSN: 0920-5691. DOI: 10.1007/s11263-007-0123-3.

[108] Lempitsky V., Rother C., Roth S., and Blake A. “Fusion Moves for Markov Random Field Optimization”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 32.8 (Aug. 2010), pp. 1392–1405. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2009.143.

[109] Lensch H. P. A., Kautz J., Goesele M., Heidrich W., and Seidel H.-P. “Image-based Reconstruction of Spatial Appearance and Geometric Detail”. In: *ACM Trans. Graph.* 22.2 (Apr. 2003), pp. 234–257. ISSN: 0730-0301. DOI: 10.1145/636886.636891.

[110] Lenzen F., Becker F., Lellmann J., Petra S., and Schnörr C. “A class of quasi-variational inequalities for adaptive image denoising and decomposition”. In: *Computational Optimization and Applications* 54.2 (2013), pp. 371–398. ISSN: 0926-6003. DOI: 10.1007/s10589-012-9456-0.

[111] Lenzen F., Kim K. I., Schäfer H., Nair R., Meister S., Becker F., Garbe C. S., and Theobalt C. “Denoising Strategies for Time-of-Flight Data”. In: *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*. Ed. by Grzegorzek M., Theobalt C., Koch R., and Kolb A. Vol. 8200. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 25–45. ISBN: 978-3-642-44963-5. DOI: 10.1007/978-3-642-44964-2_2.

[112] Lenzen F., Schäfer H., and Garbe C. “Denoising Time-Of-Flight Data with Adaptive Total Variation”. In: *Advances in Visual Computing*. Ed. by Bebis G., Boyle R., Parvin B., Koracin D., Wang S., Kyungnam K., Benes B., Moreland K., Borst C., DiVerdi S., Yi-Jen C., and Ming J. Vol. 6938. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, pp. 337–346. ISBN: 978-3-642-24027-0. DOI: 10.1007/978-3-642-24028-7_31.

[113] Levin A. and Weiss Y. “User Assisted Separation of Reflections from a Single Image Using a Sparsity Prior”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 29.9 (Sept. 2007), pp. 1647–1654. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2007.1106.

[114] Liu C., Freeman W., Adelson E., and Weiss Y. “Human-assisted motion annotation”. In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. June 2008, pp. 1–8. DOI: 10.1109/CVPR.2008.4587845.

[115] Liu S. and Cooper D. "A complete statistical inverse ray tracing approach to multi-view stereo". In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on.* June 2011, pp. 913–920. DOI: 10.1109/CVPR.2011.5995334.

[116] Lu D. and Weng Q. "A survey of image classification methods and techniques for improving classification performance". In: *International Journal of Remote Sensing* 28.5 (2007), pp. 823–870. DOI: 10.1080/01431160600746456.

[117] Marr D and Poggio T. "Cooperative computation of stereo disparity". In: *Science* 194.4262 (1976), pp. 283–287. DOI: 10.1126/science.968482. eprint: http://www.sciencemag.org/content/194/4262/283.full.pdf.

[118] Marschner S. R. and Greenberg D. P. "Inverse Lighting for Photography". In: 1. 1997, pp. 262–265. URL: http://www.ingentaconnect.com/content/ist/cic/1997/00001997/00000001/art00052.

[119] Marschner S. R., Westin S. H., Lafortune E. P., Torrance K. E., and Greenberg D. P. "Image-Based BRDF Measurement Including Human Skin". In: *Rendering Techniques '99.* Ed. by Lischinski D. and Larson G. W. Eurographics. Springer Vienna, 1999, pp. 131–144. ISBN: 978-3-211-83382-7. DOI: 10.1007/978-3-7091-6809-7_13.

[120] Matthies L. and Elfes A. "Integration of sonar and stereo range data using a grid-based representation". In: *Robotics and Automation, 1988. Proceedings., 1988 IEEE International Conference on.* Vol. 2. Apr. 1988, pp. 727–733. DOI: 10.1109/ROBOT.1988.12145.

[121] Mebius J. E. "Derivation of the Euler-Rodrigues formula for three-dimensional rotations from the general formula for four-dimensional rotations". In: *ArXiv Mathematics e-prints* (Jan. 2007). URL: http://arxiv.org/abs/math/0701759v1.

[122] Meister S. and Kondermann D. "Real versus realistically rendered scenes for optical flow evaluation". In: *Electronic Media Technology (CEMT), 2011 14th ITG Conference on.* Mar. 2011, pp. 1–6. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5936557\&tag=1.

[123] Meister S., Izadi S., Kohli P., Hämmerle M., Rother C., and Kondermann D. "When Can We Use KinectFusion for Ground Truth Acquisition?" In: (2012). URL: http://research.microsoft.com/apps/pubs/default.aspx?id=173560.

[124] Meister S., Jähne B., and Kondermann D. "Outdoor stereo camera system for the generation of real-world benchmark data sets". In: *Optical Engineering* 51.2 (2012), pp. 021107–1–021107–6. DOI: 10.1117/1.OE.51.2.021107.

[125] Meister S., Nair R., and Kondermann D. "Simulation of Time-of-Flight Sensors using Global Illumination". In: *Vision, Modeling and Visualization.* Ed. by Bronstein M., Favre J., and Hormann K. The Eurographics Association, 2013, pp. 33–40. ISBN: 978-3-905674-51-4. DOI: 10.2312/PE.VMV.VMV13.033-040.

[126] Meister S. N. R. "On Creating Reference Data for Performance Analysis in Image Processing". dissertation. 2014. URL: http://archiv.ub.uni-heidelberg.de/volltextserver/16193/.

[127] Moharam M. G. and Gaylord T. K. “Rigorous coupled-wave analysis of planar-grating diffraction”. In: *J. Opt. Soc. Am.* 71.7 (July 1981), pp. 811–818. DOI: 10.1364/JOSA.71.000811.

[128] Morales S. and Klette R. “A Third Eye for Performance Evaluation in Stereo Sequence Analysis”. In: *Computer Analysis of Images and Patterns*. Ed. by Jiang X. and Petkov N. Vol. 5702. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2009, pp. 1078–1086. ISBN: 978-3-642-03766-5. DOI: 10.1007/978-3-642-03767-2_131.

[129] Müller G., Meseth J., Sattler M., Sarlette R., and Klein R. “Acquisition, Synthesis, and Rendering of Bidirectional Texture Functions”. In: vol. 24. 1. Blackwell Publishing Ltd., 2005, pp. 83–109. DOI: 10.1111/j.1467-8659.2005.00830.x.

[130] Murphy K. P., Weiss Y., and Jordan M. I. “Loopy Belief Propagation for Approximate Inference: An Empirical Study”. In: *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*. UAI’99. Stockholm, Sweden: Morgan Kaufmann Publishers Inc., 1999, pp. 467–475. ISBN: 1-55860-614-9. URL: http://dl.acm.org/citation.cfm?id=2073796.2073849.

[131] Nair R., Lenzen F., Meister S., Schäfer H., Garbe C., and Kondermann D. “High Accuracy TOF and Stereo Sensor Fusion at Interactive Rates”. In: *ECCV 2012. Workshops and Demonstrations*. Ed. by Fusiello A., Murino V., and Cucchiara R. Vol. 7584. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, pp. 1–11. ISBN: 978-3-642-33867-0. DOI: 10.1007/978-3-642-33868-7_1.

[132] Nair R., Meister S., Lambers M., Balda M., Hofmann H., Kolb A., Kondermann D., and Jähne B. “Ground Truth for Evaluating Time of Flight Imaging”. In: *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*. Ed. by Grzegorzek M., Theobalt C., Koch R., and Kolb A. Vol. 8200. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 52–74. ISBN: 978-3-642-44963-5. DOI: 10.1007/978-3-642-44964-2_4.

[133] Nair R., Ruhl K., Lenzen F., Meister S., Schäfer H., Garbe C. S., Eisemann M., Magnor M., and Kondermann D. “A Survey on Time-of-Flight Stereo Fusion”. In: *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*. Ed. by Grzegorzek M., Theobalt C., Koch R., and Kolb A. Vol. 8200. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 105–127. ISBN: 978-3-642-44963-5. DOI: 10.1007/978-3-642-44964-2_6.

[134] Nemhauser G. L. and Wolsey L. A. *Integer and Combinatorial Optimization*. New York, NY, USA: Wiley-Interscience, 1988. ISBN: 0-471-82819-X. URL: http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0471359432.html.

[135] Newcombe R. A., Izadi S., Hilliges O., Molyneaux D., Kim D., Davison A. J., Kohi P., Shotton J., Hodges S., and Fitzgibbon A. “KinectFusion: Real-time dense surface mapping and tracking”. In: *Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on*. Oct. 2011, pp. 127–136. DOI: 10.1109/ISMAR.2011.6092378.

[136] Nocedal J. and Wright S. J. *Numerical Optimization 2nd Edition.* Springer, 2006.

[137] Onkarappa N. and Sappa A. D. "Synthetic sequences and ground-truth flow field generation for algorithm validation". In: *Multimedia Tools and Applications* (2013), pp. 1–15. ISSN: 1380-7501. DOI: `10.1007/s11042-013-1771-7`.

[138] Pal N. R. and Pal S. K. "A review on image segmentation techniques". In: *Pattern Recognition* 26.9 (1993), pp. 1277 –1294. ISSN: 0031-3203. DOI: `http://dx.doi.org/10.1016/0031-3203(93)90135-J`.

[139] Papenberg N., Bruhn A., Brox T., Didas S., and Weickert J. "Highly Accurate Optic Flow Computation with Theoretically Justified Warping". In: *International Journal of Computer Vision* 67.2 (2006), pp. 141–158. ISSN: 0920-5691. DOI: `10.1007/s11263-005-3960-y`.

[140] Park J. B. "Detection of Specular Highlights in Color Images using a New Color Space Transformation". In: *Robotics and Biomimetics, 2004. ROBIO 2004. IEEE International Conference on.* Aug. 2004, pp. 737–741. DOI: `10.1109/ROBIO.2004.1521873`.

[141] Park J., Kim H., Tai Y.-W., Brown M., and Kweon I. "High quality depth map upsampling for 3D-TOF cameras". In: *Computer Vision (ICCV), 2011 IEEE International Conference on.* Nov. 2011, pp. 1623–1630. DOI: `10.1109/ICCV.2011.6126423`.

[142] Patow G. and Pueyo X. "A Survey of Inverse Rendering Problems". In: vol. 22. 4. Blackwell Publishing, Inc, 2003, pp. 663–687. DOI: `10.1111/j.1467-8659.2003.00716.x`.

[143] Payne A. D., Jongenelen A. P., Dorrington A. A., Cree M. J., and Carnegie D. A. "Multiple frequency range imaging to remove measurement ambiguity". In: *Proceedings of the 9th Conference on Optical 3-D Measurement Techniques.* 2009, pp. 139–148. URL: `http://researchcommons.waikato.ac.nz/handle/10289/4032`.

[144] Perona P. and Malik J. "Scale-space and edge detection using anisotropic diffusion". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 12.7 (July 1990), pp. 629–639. ISSN: 0162-8828. DOI: `10.1109/34.56205`.

[145] Pharr M. and Humphreys G. *Physically Based Rendering: From Theory to Implementation.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2004. ISBN: 012553180X. URL: `http://www.pbrt.org/`.

[146] Poggio T., Torre V., and Koch C. "Computational Vision and Regularization Theory". In: *Nature* 317.26 (Sept. 1985), pp. 314–319. ISSN: 0028-0836. DOI: `10.1038/317314a0`.

[147] Poppinga J. and Birk A. "A Novel Approach to Efficient Error Correction for the Swiss-Ranger Time-of-Flight 3D Camera". In: *RoboCup 2008: Robot Soccer World Cup XII.* Ed. by Iocchi L., Matsubara H., Weitzenfeld A., and Zhou C. Vol. 5399. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2009, pp. 247–258. ISBN: 978-3-642-02920-2. DOI: `10.1007/978-3-642-02921-9_22`.

[148] Prados E. and Faugeras O. "Shape From Shading". In: *Handbook of Mathematical Models in Computer Vision.* Ed. by Paragios N., Chen Y., and Faugeras O. Springer US, 2006, pp. 375–388. ISBN: 978-0-387-26371-7. DOI: `10.1007/0-387-28831-7_23`.

[149] Rajagopalan A., Chaudhuri S., and Mudenagudi U. “Depth estimation and image restoration using defocused stereo pairs”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 26.11 (Nov. 2004), pp. 1521–1525. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2004.102.

[150] Rall L. B. *Automatic Differentiation: Techniques and Applications.* Vol. 120. Lecture Notes in Computer Science. Berlin: Springer, 1981. DOI: 10.1007/3-540-10861-0.

[151] Reynolds M., Dobos J., Peel L., Weyrich T., and Brostow G. “Capturing Time-of-Flight data with confidence”. In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on.* June 2011, pp. 945–952. DOI: 10.1109/CVPR.2011.5995550.

[152] Rhemann C., Hosni A., Bleyer M., Rother C., and Gelautz M. “Fast cost-volume filtering for visual correspondence and beyond”. In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on.* June 2011, pp. 3017–3024. DOI: 10.1109/CVPR.2011.5995372.

[153] Roth S. and Black M. “Specular Flow and the Recovery of Surface Structure”. In: *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on.* Vol. 2. 2006, pp. 1869–1876. DOI: 10.1109/CVPR.2006.290.

[154] Rudin L. I., Osher S., and Fatemi E. “Nonlinear total variation based noise removal algorithms”. In: *Physica D: Nonlinear Phenomena* 60.1 (1992), pp. 259 –268. ISSN: 0167-2789. DOI: http://dx.doi.org/10.1016/0167-2789(92)90242-F.

[155] Ruhl K., Klose F., Lipski C., and Magnor M. “Integrating Approximate Depth Data into Dense Image Correspondence Estimation”. In: *Proceedings of the 9th European Conference on Visual Media Production.* CVMP ’12. London, United Kingdom: ACM, 2012, pp. 26–31. ISBN: 978-1-4503-1311-7. DOI: 10.1145/2414688.2414692.

[156] Schäfer H., Lenzen F., and Garbe C. “Depth and Intensity Based Edge Detection in Time-of-Flight Images”. In: *3D Vision - 3DV 2013, 2013 International Conference on.* June 2013, pp. 111–118. DOI: 10.1109/3DV.2013.23.

[157] Schäfer H. “Image Enhancement and Parameter Estimation for Time-of-Flight Cameras”. dissertation. 2014. URL: http://archiv.ub.uni-heidelberg.de/volltextserver/17760/.

[158] Scharstein D. and Szeliski R. “High-accuracy stereo depth maps using structured light”. In: *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on.* Vol. 1. June 2003, pp. 195–202. DOI: 10.1109/CVPR.2003.1211354.

[159] Scharstein D. and Szeliski R. “A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms”. In: *International Journal of Computer Vision* 47.1-3 (2002), pp. 7–42. ISSN: 0920-5691. DOI: 10.1023/A:1014573219977.

[160] Scharstein D. and Szeliski R. “Stereo Matching with Nonlinear Diffusion”. In: *International Journal of Computer Vision* 28.2 (1998), pp. 155–174. ISSN: 0920-5691. DOI: 10.1023/A:1008015117424.

[161] Schechner Y. Y. and Kiryati N. “Depth from Defocus vs. Stereo: How Different Really Are They?” In: *International Journal of Computer Vision* 39.2 (2000), pp. 141–162. ISSN: 0920-5691. DOI: 10.1023/A:1008175127327.

[162] Schiller I., Beder C., and Koch R. “Calibration of a PMD-camera using a planar calibration pattern together with a multi-camera setup”. In: *The international archives of the photogrammetry, remote sensing and spatial information sciences* 37 (2008), pp. 297–302. URL: http://www.isprs.org/proceedings/XXXVII/congress/3_pdf/46.pdf.

[163] Schmidt M. “Analysis, modeling and dynamic optimization of 3D time-of-flight imaging systems”. Dissertation. IWR, Faculty of Physics, Univ. Heidelberg, 2011. URL: http://www.ub.uni-heidelberg.de/archiv/12297.

[164] Schwarte R., Xu Z., Heinol H.-G., Olk J., Klein R., Buxbaum B., Fischer H., and Schulte J. “New electro-optical mixing and correlating sensor: facilities and applications of the photonic mixer device (PMD)”. In: vol. 3100. 1997, pp. 245–253. DOI: 10.1117/12.287751.

[165] Schwartz C., Sarlette R., Weinmann M., and Klein R. “DOME II: A Parallelized BTF Acquisition System”. In: *Proceedings of the Eurographics 2013 Workshop on Material Appearance Modeling: Issues and Acquisition*. MAM ’13. Zaragoza, Spain: Eurographics Association, 2013, pp. 25–31. ISBN: 978-3-905674-48-4. DOI: 10.2312/MAM.MAM2013.025-031.

[166] Seitz S., Curless B., Diebel J., Scharstein D., and Szeliski R. “A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms”. In: *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. Vol. 1. June 2006, pp. 519–528. DOI: 10.1109/CVPR.2006.19.

[167] Shi J. and Malik J. “Normalized cuts and image segmentation”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22.8 (Aug. 2000), pp. 888–905. ISSN: 0162-8828. DOI: 10.1109/34.868688.

[168] Smith A., Skorupski J., and Davis J. *Transient Rendering*. Tech. rep. UCSC-SOE-08-26. School of Engineering, University of California, Santa Cruz, Feb. 2008. URL: http://classes.soe.ucsc.edu/cmps290b/Fall07/TransientRendering/.

[169] Song Y., Glasbey C. A., Heijden G. W. van der, Polder G., and Dieleman J. A. “Combining Stereo and Time-of-Flight Images with Application to Automatic Plant Phenotyping”. In: *Image Analysis*. Ed. by Heyden A. and Kahl F. Vol. 6688. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, pp. 467–478. ISBN: 978-3-642-21226-0. DOI: 10.1007/978-3-642-21227-7_44.

[170] Sperling G. “Binocular vision: A physical and a neural theory”. In: *The American Journal of Psychology* (1970), pp. 461–534. DOI: 10.2307/1420686.

[171] Spielmann S., Helzle V., and Nair R. “On-set Depth Capturing for VFX Productions Using Time of Flight”. In: *ACM SIGGRAPH 2013 Talks*. SIGGRAPH ’13. Anaheim, California: ACM, 2013, 13:1–13:1. ISBN: 978-1-4503-2344-4. DOI: 10.1145/2504459.2504475.

[172] Spirig T., Seitz P., Vietze O., and Heitger F. “The lock-in CCD-two-dimensional synchronous detection of light”. In: *Quantum Electronics, IEEE Journal of* 31.9 (Sept. 1995), pp. 1705–1708. ISSN: 0018-9197. DOI: 10.1109/3.406386.

[173] Strecha C., von Hansen W., Van Gool L., Fua P., and Thoennessen U. “On benchmarking camera calibration and multi-view stereo for high resolution imagery”. In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.* June 2008, pp. 1–8. DOI: 10.1109/CVPR.2008.4587706.

[174] Szeliski R., Zabih R., Scharstein D., Veksler O., Kolmogorov V., Agarwala A., Tappen M., and Rother C. “A Comparative Study of Energy Minimization Methods for Markov Random Fields with Smoothness-Based Priors”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 30.6 (June 2008), pp. 1068–1080. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2007.70844.

[175] Takeda Y., Hiura S., and Sato K. “Fusing Depth from Defocus and Stereo with Coded Apertures”. In: *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on.* June 2013, pp. 209–216. DOI: 10.1109/CVPR.2013.34.

[176] Taylor J., Shotton J., Sharp T., and Fitzgibbon A. “The Vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation”. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.* June 2012, pp. 103–110. DOI: 10.1109/CVPR.2012.6247664.

[177] Theobalt C., Grzegorzek M., Koch R., and Kolb A. *Time-of-Flight and Depth Imaging.* Springer, 2012. DOI: 10.1007/978-3-642-44964-2.

[178] Treuille A., Hertzmann A., and Seitz S. M. “Example-Based Stereo with General BRDFs”. In: *Computer Vision - ECCV 2004.* Ed. by Pajdla T. and Matas J. Vol. 3022. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2004, pp. 457–469. ISBN: 978-3-540-21983-5. DOI: 10.1007/978-3-540-24671-8_36.

[179] Triggs B., McLauchlan P. F., Hartley R. I., and Fitzgibbon A. W. “Bundle Adjustment: A Modern Synthesis”. In: *Vision Algorithms: Theory and Practice.* Ed. by Triggs B., Zisserman A., and Szeliski R. Vol. 1883. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2000, pp. 298–372. ISBN: 978-3-540-67973-8. DOI: 10.1007/3-540-44480-7_21.

[180] Tsin Y., Kang S. B., and Szeliski R. “Stereo matching with reflections and translucency”. In: *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on.* Vol. 1. June 2003, pp. 702–709. DOI: 10.1109/CVPR.2003.1211422.

[181] Vaudrey T., Rabe C., Klette R., and Milburn J. “Differences between stereo and motion behaviour on synthetic and real-world stereo sequences”. In: *Image and Vision Computing New Zealand, 2008. IVCNZ 2008. 23rd International Conference.* Nov. 2008, pp. 1–6. DOI: 10.1109/IVCNZ.2008.4762133.

[182] Viola P. and Jones M. “Rapid object detection using a boosted cascade of simple features”. In: *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on.* Vol. 1. 2001, pp. 511–518. DOI: 10.1109/CVPR.2001.990517.

[183] Wanner S. and Goldluecke B. “Reconstructing Reflective and Transparent Surfaces from Epipolar Plane Images”. In: *Pattern Recognition.* Ed. by Weickert J., Hein M., and Schiele B. Vol. 8142. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 1–10. ISBN: 978-3-642-40601-0. DOI: 10.1007/978-3-642-40602-7_1.

[184] Ward G. J. “Measuring and Modeling Anisotropic Reflection”. In: *SIGGRAPH Comput. Graph.* 26.2 (July 1992), pp. 265–272. ISSN: 0097-8930. DOI: 10.1145/142920.134078.

[185] Wedel A., Pock T., Zach C., Bischof H., and Cremers D. “An Improved Algorithm for TV-L 1 Optical Flow”. In: *Statistical and Geometrical Approaches to Visual Motion Analysis.* Ed. by Cremers D., Rosenhahn B., Yuille A., and Schmidt F. Vol. 5604. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2009, pp. 23–45. ISBN: 978-3-642-03060-4. DOI: 10.1007/978-3-642-03061-1_2.

[186] Woodford O., Torr P., Reid I., and Fitzgibbon A. “Global Stereo Reconstruction under Second-Order Smoothness Priors”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31.12 (Dec. 2009), pp. 2115–2128. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2009.131.

[187] Wulff J., Butler D. J., Stanley G. B., and Black M. J. “Lessons and Insights from Creating a Synthetic Optical Flow Benchmark”. In: *Computer Vision , ECCV 2012. Workshops and Demonstrations.* Ed. by Fusiello A., Murino V., and Cucchiara R. Vol. 7584. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, pp. 168–177. ISBN: 978-3-642-33867-0. DOI: 10.1007/978-3-642-33868-7_17.

[188] Yang Q., Tan K.-H., Culbertson B., and Apostolopoulos J. “Fusion of active and passive sensors for fast 3D capture”. In: *Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on.* Oct. 2010, pp. 69–74. DOI: 10.1109/MMSP.2010.5661996.

[189] Yu Y., Debevec P., Malik J., and Hawkins T. “Inverse Global Illumination: Recovering Reflectance Models of Real Scenes from Photographs”. In: *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques.* SIGGRAPH ’99. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1999, pp. 215–224. ISBN: 0-201-48560-5. DOI: 10.1145/311535.311559.

[190] Zach C., Pock T., and Bischof H. “A Duality Based Approach for Realtime TV-L 1 Optical Flow”. In: *Pattern Recognition.* Ed. by Hamprecht F., Schnörr C., and Jähne B. Vol. 4713. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2007, pp. 214–223. ISBN: 978-3-540-74933-2. DOI: 10.1007/978-3-540-74936-3_22.

[191] Zhang Z. “A flexible new technique for camera calibration”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22.11 (Nov. 2000), pp. 1330–1334. ISSN: 0162-8828. DOI: 10.1109/34.888718.

[192] Zhu J., Wang L., Gao J., and Yang R. “Spatial-Temporal Fusion for High Accuracy Depth Maps Using Dynamic MRFs”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 32.5 (May 2010), pp. 899–909. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2009.68.

[193] Zhu J., Wang L., Yang R., and Davis J. “Fusion of time-of-flight depth and stereo for high accuracy depth maps”. In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.* June 2008, pp. 1–8. DOI: 10.1109/CVPR.2008.4587761.

[194] Zhu J., Wang L., Yang R., Davis J. E., and Pan Z. “Reliability Fusion of Time-of-Flight Depth and Stereo Geometry for High Quality Depth Maps”. In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 33.7 (July 2011), pp. 1400–1414. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2010.172.

[195] Zisserman A., Giblin P., and Blake A. “The information available to a moving observer from specularities”. In: *Image and Vision Computing* 7.1 (1989), pp. 38 –42. ISSN: 0262-8856. DOI: http://dx.doi.org/10.1016/0262-8856(89)90018-8.