Multi-Scale Surface Reconstruction from Images

Vom Fachbereich Informatik
der Technischen Universität Darmstadt
zur Erlangung des akademischen Grades
Dr. rer. nat.
genehmigte

DISSERTATION

von

Dipl.-Math. Ronny Klowsky
geb. in Jena, Deutschland

Referenten der Arbeit:	Prof. Dr.-Ing. Michael Goesele Technische Universität Darmstadt Prof. Dr. Leif Kobbelt Rheinisch-Westfälische Technische Hochschule Aachen
Tag der Einreichung:	14/08/2013
Tag der mündlichen Prüfung:	30/09/2013
Erscheinungsjahr:	2014

Darmstädter Dissertation
D 17

Erklärung zur Dissertation

Hiermit versichere ich, die vorliegende Dissertation selbständig nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Darmstadt, den 14.08.2013

Ronny Klowsky

Abstract

Many surface reconstruction algorithms have been developed to process point data originating from laser scans. Because laser scanning is a very expensive technique and not available to everyone, 3D reconstruction from images (using, *e.g.*, *multi-view stereo*) is a promising alternative. In recent years a lot of progress has been made in the computer vision domain and nowadays algorithms are capable of reconstructing large 3D scenes from consumer photographs. Whereas laser scans are very controlled and typically only a few scans are taken, images may be subject to more uncontrolled variations. Standard multi-view stereo algorithms give rise to multi-scale data points due to different camera resolutions, focal lengths, or various distances to the object. When reconstructing a surface from this data, the multi-scale property has to be taken into account because the assumption that the points are samples from the true surface might be violated.

This thesis presents two surface reconstruction algorithms that take resolution and scale differences into account. In the first approach we model the uncertainty of each sample point according to its *footprint*, the surface area that was taken into account during multi-view stereo. With an adaptive volumetric resolution, also steered by the footprints of the sample points, we achieve detailed reconstructions even for large-scale scenes. Then, a general wavelet-based surface reconstruction framework is presented. The multi-scale sample points are characterized by a convolution kernel and the points are fused in frequency space while preserving locality. We suggest a specific implementation for 2.5D surfaces that incorporates our theoretic findings about sample points originating from multi-view stereo and shows promising results on real-world data sets.

The other part of the thesis analyzes the scale characteristics of patch-based depth reconstruction as used in many (multi-view) stereo techniques. It is driven by the question how the reconstruction preserves surface details or high frequencies. We introduce an intuitive model for the reconstruction process, prove that it yields a linear system and determine the modulation transfer function. This allows us to predict the amplitude loss of high frequencies in connection with the used patch-size and the internal and external camera parameters. Experiments on synthetic and real-world data demonstrate the ac-

curacy of our model but also show the limitations. Finally, we propose a generalization of the model allowing for weighted patch fitting. The reconstructed points can then be described by a convolution of the original surface and we show how weighting the pixels during photo-consistency optimization affects the smoothing kernel. In this way we are able to connect a standard notion of smoothing to multi-view stereo reconstruction.

In summary, this thesis provides a profound analysis of patch-based (multi-view) stereo reconstruction and introduces new concepts for surface reconstruction from the resulting multi-scale sample points.

Zusammenfassung

Viele Oberflächenrekonstruktions-Algorithmen wurden für Punktdaten entwickelt, die bei der Verwendung von Laserscannern entstehen. Da die Technik des Laserscannings sehr teuer und nicht für jedermann verfügbar ist, erscheint die 3D-Rekonstruktion aus Bildern als eine vielversprechende Alternative. In den letzten Jahren konnten auf dem Gebiet der Computer Vision viele Fortschritte erzielt werden und heutige Algorithmen sind in der Lage, große Szenen aus Fotos von Normalverbrauchern zu rekonstruieren. Während Laserscans sehr gezielt durchgeführt werden und typischerweise nur wenige Aufnahmen notwendig sind, können Bilder sehr viel unterschiedlicher sein. Verschiedene Bildauflösungen, Brennweiten oder Entfernungen zum Objekt führen mit üblichen Multi-view Stereo Methoden zu Punkten mit multiplen Skalen. Ein Algorithmus zur Oberflächenrekonstruktion aus diesen Daten sollte die verschiedenen Skalen berücksichtigen, denn die übliche Annahme, dass die Punkte von der unbekannten Oberfläche gesampelt sind, könnte verletzt sein.

In dieser Arbeit werden zwei neue Algorithmen zur Oberflächenrekonstruktion vorgestellt, die Unterschiede in der Auflösung und verschiedene Skalen mit einbeziehen. Der erste Ansatz modelliert die Ungenauigkeit der Punkte in Abhängigkeit von ihrem *Footprint*, das ist der Teil der Oberfläche der zur Rekonstruktion dieses Punktes durch Multi-view Stereo in Betracht gezogen wurde. Durch eine adaptive räumliche Auflösung, die ebenfalls durch den Footprint gesteuert wird, erzielen wir auch für große Szenen detaillierte Rekonstruktionen. Als Zweites wird ein Wavelet-basiertes Framework zur Oberflächenrekonstruktion vorgestellt. Die Punkte auf multiplen Skalen werden durch Faltungskernel charakterisiert und im Frequenzraum vereinigt, wobei die Lokalität beachtet wird. Wir stellen eine konkrete Implementierung für 2,5D Oberflächen vor, die unsere theoretischen Erkenntnisse über Multi-view Stereo Punkte einbezieht und vielversprechende Ergebnisse auf realen Daten erzielt.

Der andere Teil dieser Dissertation analysiert die Skalen-Charakteristik von Patch-basierter Tiefenrekonstruktion, wie sie von Multi-view Stereo Verfahren verwendet wird. Wir gehen dabei der Frage nach, inwieweit Oberflächendetails oder hohe Frequenzen

durch die Multi-view Stereo Rekonstruktion erhalten bleiben. Wir verwenden dazu ein intuitives Modell, das den Rekonstruktionsprozess abbildet, weisen nach, dass es sich um ein lineares System handelt und bestimmen die Modulationsübertragungsfunktion. Diese erlaubt uns vorherzusagen, wie sich die Amplitude von hohen Frequenzen in Abhängigkeit von der verwendeten Patchgröße und den externen und internen Kameraparametern verringert. Experimente auf synthetischen und realen Daten demonstrieren die Genauigkeit unseres Modells, zeigen aber auch die Grenzen auf. Wir erweitern anschließend das Modell, um auch gewichtetes Patch Fitting abbilden zu können. Die rekonstruierten Punkte können mithilfe einer Faltung der ursprünglichen Oberfläche beschrieben werden und wir zeigen den Zusammenhang zwischen der gewichteten Photokonsistenz-Optimierung und dem Filterkern. Damit verknüpfen wir die Multi-Skalen Rekonstruktion mit der üblichen Vorstellung einer Glättung.

Die vorgelegte Arbeit enthält damit eine fundierte Analyse von Patch-basierten (Multi-View) Stereo Rekonstruktionsverfahren und offeriert neue Konzepte zur Oberflächenrekonstruktion aus den resultierenden Multi-Skalen Punktdaten.

Contents

1 Introduction

Contents

THE need for 3D models of real-world objects or scenes arises in various fields such as games or movies, medical applications, natural sciences such as geology, or in the context of cultural heritage. For a long time only active technologies such as laser scanning or structured light scanning allowed to reliably capture the 3D geometry of an object. These techniques were primarily used to digitize tools or prototypes in the computer-aided design process (*reverse engineering*). Active scanning technology is still further developed and yields very accurate point clouds. It is widely used, *e.g.*, to digitize cultural heritage sculptures and architecture such as in the Digital Michelangelo Project[1]. Active scanning devices have some drawbacks, too, mainly that the devices are still expensive and not widespread. Passive scanning technology just relying on photographs is a promising alternative and allows also non-expert users to capture 3D objects or even larger scenes. It is very easy to capture a set of images from a certain object or scene since many people own a camera (or nowadays a mobile phone with an integrated camera). Additionally, there are a lot of images available on the Internet where users upload their photographs to sharing platforms such as flickr[2].

The next section shortly describes the development of multi-view stereo methods in recent years building the bridge to multi-scale sample points. Section 1.3 then summarizes the main thesis contributions.

[1] http://graphics.stanford.edu/data/mich/

[2] http://flickr.com

1.1 Multi-View Stereo

The introduction of robust feature matching (*e.g.*, SIFT [Lowe 2004]) lead to structure-from-motion algorithms that are able to register images even if the data is uncontrolled such as downloaded images from community photo collections [Snavely et al. 2006, Snavely et al. 2008]. So far, multi-view stereo algorithms were mainly applied to images taken under controlled, laboratory conditions but now the application on real-world scenes came into focus [Goesele et al. 2007, Furukawa and Ponce 2010]. At the same time growing capabilities in computation, post-processing, and rendering, led to the desire to capture all kinds of objects such as statues, buildings, places, or even entire cities (not only in 3D but with changes over time, *e.g.*, in the the 4D cities project[3]). Thus the focus of newly developed algorithms shifted towards scalability to allow reconstructions of large scenes [Labatut et al. 2007, Jancosek et al. 2009, Hiep et al. 2009, Furukawa et al. 2010]. A famous example is the "Building rome in a day" project [Agarwal et al. 2009, Frahm et al. 2010]. Of course these techniques typically work fully automatic and are even robust enough to provide a web service[4] where non-professional users can upload images and obtain a 3D reconstruction within at most a few hours.

Ideally, even when reconstructing large scenes the 3D model still reflects a high level of detail reproducing small surface variations, edges, and corners. It has been shown in the famous Middlebury benchmark[5] that multi-view stereo reconstructions on controlled data have the potential to very accurately recover surface details [Seitz et al. 2006]. Another benchmark took this further to outdoor scenes and also here reconstructions show a remarkable accuracy [Strecha et al. 2008]. Looking at the result of Goesele *et al.* [2007] where they compared the reconstruction of the Pisa model with a laser scan it is even indicated that multi-view stereo methods achieve as accurate reconstruction results as classical laser scanners. We therefore think it is worth to further investigate the topic of 3D surface reconstruction from images and push the technology to the next level.

When looking at multi-view stereo algorithms that have proven to scale to large scenes basically all reconstruct a point cloud, at least as an intermediate step. If a closed surface is desired, such as a triangle mesh, an implicit function or an explicit parameterization, a surface reconstruction algorithm is applied that takes the reconstructed point cloud as input. Reconstructing a surface from a point cloud is a well-researched topic but still far from being solved. These algorithms were originally developed to process points

[3] http://www.cc.gatech.edu/4d-cities/dhtml/index.html
[4] http://www.arc3d.be
[5] http://vision.middlebury.edu/mview/

originating from laser scanners, typically taking a few scans to cover the entire object. Although some of these methods were successfully applied to multi-view stereo points, this input data has quite different characteristics. Sample points from multi-view stereo are typically less complete and accurate, vary spatially in resolution but most importantly, samples might emerge from multiple scales.

1.2 Multi-Scale Sample Points

The basis of practically all multi-view stereo algorithms is to find corresponding image positions in several views. Correspondences are often determined using a patch-based photo-consistency measure. Popular choices are the normalized cross-correlation (NCC) or the sum of squared distances (SSD). In order to compute these photo-consistency measures a small planar patch in 3D is projected into the images and sampled at a predefined number of positions. The size of the patch is usually defined by the projected size in the images or in one particular image, often called reference image, in order to obtain meaningful sampling distances of approximately the size of the pixel spacing. Consequently, the real world patch size depends on the image resolution, focal length of the camera, and its distance to object. If the orientation of the patch is varied throughout the matching process then the patch size also depends on the surface normals. Throughout this thesis we use the term *fine scale* sample to refer to sample points that were reconstructed using a small patch in contrast to *coarse scale* samples where the underlying patch is of larger size.

The main focus of this thesis is how coarse and fine scale should be handled in a surface reconstruction algorithm. The insight that fine scale sample points have the potential to capture surface details whereas coarse scale points reflect more the base structure guides one of the presented approaches to surface reconstruction in this thesis. We continue by deeper analyzing the difference between coarse and fine scale samples and model the fitting process of standard patch-based multi-view stereo algorithms. In particular, we will show that the reconstructed sample points do not necessarily lie on the true surface but on smoothed versions of the true surface. This contradicts the widely used paradigm in surface reconstruction that the input are real point samples with mean position on the true surface and emphasizes even more that it is necessary to consider the multi-scale property of the sample points in a surface reconstruction approach. In order to do so properly the smoothing has to be characterized in a mathematical way providing a *generative model*. Among others we will show that the commonly used convolution operator can appropriately describe the smoothing that occurs in a multi-view stereo

reconstruction.

1.3 Contributions of the Thesis

The contributions of this thesis affect two major research areas. The first is surface reconstruction from sample points. Algorithms exist that cope with multi-resolution sample points, *i.e.*, spatially varying sampling distributions, and using multi-resolution data structures to support different reconstruction resolutions. However, to our knowledge we are the first to specifically model and handle multi-scale input data. The second area is multi-view stereo where we present an analysis of patch-based depth reconstruction with the focus on how surface details are preserved depending on the patch size. We model the systematic error in the reconstruction and provide the means to achieve better frequency behavior using a weighted matching scheme.

1.3.1 Contributions to Surface Reconstruction

This thesis presents two new algorithms with different ways to handle multi-scale input data. In the first algorithm we argue that many measurement techniques actually take a small surface area into account to acquire a sample point. We refer to that area as the *footprint* of a sample and take it, or an estimate, into account during the reconstruction process. The intuition is that sample points with a small footprint can capture surface details far better than sample points with a large footprint. We integrate this intuition into an existing robust surface reconstruction algorithm that creates a confidence map in 3D space. Each sample point adds a confidence distribution that depends on its footprint. Additionally, we extend the existing method to be applicable to a very general class of input data with arbitrary surface shape and genus. The footprints of the sample points also influence the local volumetric resolution we use for building the confidence volume and thus for surface reconstruction. This allows us to reconstruct fine details exactly and only at locations where fine scale input samples are available.

The second proposed algorithm reconstructs a 2.5D height field surface by fusing the multi-scale sample points in frequency space. A wavelet decomposition allows for operation in only those space-frequency windows that are influenced by the individual sample points. The associated scales of the sample points, in terms of the convolution kernels, also steer the detail level in the final reconstruction. The algorithmic framework is applicable to all multi-scale data that can be characterized by (an estimated) convolution kernel. To our knowledge, this is the first approach to combine coarse- and fine-scale point data taking into account that the sample points do not lie on the true surface.

1.3.2 Contributions to Multi-View Stereo

Multi-view stereo methods often use patch-based matching in order to determine the 3D position of a point or the depth of a pixel. In choosing the size of the patch common knowledge is that there is a trade-off between accuracy and robustness. This thesis explores the influence of the matching window giving new insight into a broad range of multi-view stereo algorithms. We propose to model the patch-based depth reconstruction by fitting a planar patch in the least squares sense to the (unknown) true surface. This corresponds to a widely spread intuition of patch-based depth reconstruction using photo-consistency measures. Under this assumption we prove that the reconstruction process fulfills the linear system requirements and determine the modulation transfer function. Experiments on synthetic as well as real-world data sets show that our model convincingly captures the behavior of a popular multi-view stereo algorithm. As a result, there is a significant amplitude loss in the multi-view stereo depth reconstruction depending on the details in the unknown surface (frequencies) and the reconstruction resolution. With our theoretical model we can predict this reconstruction error. Furthermore, we can correct the amplitude of fine scale details in the reconstruction accordingly within the limits of the imperfect reconstruction.

In a second step we the turn the theoretical analysis from Fourier space to geometry space. This allows us to express the reconstructed surface in terms of a convolution of the original surface with some kernel. Thus, recovering the original surface is similar to a deconvolution problem and not well-posed. Using standard matching the pixels in a patch are weighted equally and the convolution kernel is a box filter. The thesis then establishes the connection between weighted (multi-view) stereo depth reconstruction and the resulting geometry. We show that under certain assumptions the convolution kernel is a dilated version of the weighting function. For example, using Gaussian weighting results in nicely low-pass filtered reconstructions instead of the occurring high-frequency artifacts when using uniform weights. This is again experimentally validated on synthetic and real-world data.

1.4 Thesis Overview and Structure

In the following we give an overview over the structure of the thesis and a short summary of each chapter. The ordering of the chapters follows the line of the development with increasing insight. At the same time, this reflects the chronological order of the individual research projects and the corresponding publications.

Chapter 2 gives a very general overview of work that is related to this thesis. It cov-

ers multi-view stereo methods and surface reconstruction from images as well surface reconstruction from point clouds in general. Each following chapter provides an additional, more specific view on related work closely related to the presented research.

Before going into detail, **Chapter 3** introduces the general pipeline used throughout this thesis to reconstruct a surface from images. The individual steps are shortly described and linked to the content of the thesis.

Chapter 4 presents a new hierarchical surface reconstruction approach exploiting the footprint information which is inherent to each sample point. This work started with the Masters thesis by Patrick Mücke under the supervision of the thesis author. The corresponding publication "Surface reconstruction from multi-resolution sample points" [Mücke et al. 2011] won the best paper award at VMV. The further development of the method was published under the title "Hierarchical Surface Reconstruction from Multi-resolution Point Samples" [Klowsky et al. 2012b] in the Springer LNCS series which corresponds to the content of the chapter. The source code is solely written by Patrick Mücke and available on the project website [Mücke et al. 2012].

In **Chapter 5** we analyze patch-based depth reconstruction theoretically and show that it can be modeled as a linear system. We determine the modulation transfer function to be a sinc which corresponds to a convolution with a box filter. We validate this experimentally on synthetic as well as real-world data. This chapter corresponds to the paper "Modulation transfer function of patch-based stereo systems" [Klowsky et al. 2012a] presented at CVPR.

Using a weighted patch-based stereo we show a generalization of the model in **Chapter 6**. A broad range of convolution filters can be realized. We determine necessary criteria that the weighting function has to fulfill. As a special case, using Gaussian weighting with different standard deviations reconstructs a scale-space representation of the original surface. This work was published at SSVM as "Weighted patch-based reconstruction: linking (multi-view) stereo to scale space" [Klowsky et al. 2013].

Chapter 7 presents a reconstruction framework for 2.5D height field surfaces where the sample points are fused in frequency space. This algorithm is ideally suited to process data created with the weighted patch-based depth reconstruction from the previous chapter. This content of this chapter has been published as a technical report [Klowsky and Goesele 2013].

Finally, **Chapter 8** summarizes the contributions of the thesis and concludes with an outlook on future work.

2 Related Work

THIS chapter gives a broad overview of prior work covering research areas touched on by this thesis. Each following chapter of the thesis provides a more detailed discussion of particularly relevant related work and discusses the distinction compared to our work. In the following we distinguish between geometry reconstruction from images and surface reconstruction from sample points. There is no tight boundary though because quite a few methods from the first category create sample points in the first place and then apply standard surface reconstruction algorithms. On the other hand, classical surface reconstruction algorithms have initially been designed to process accurate, densely sampled data points, *e.g.*, from laser scanners.

2.1 Geometry Reconstruction from Images

2.1.1 Volumetric Representation

Using a volumetric representation is probably pioneered by Seitz and Dyer [1999]. They propose *voxel coloring* where they compute the visibility and color for each voxel using a multi-view photo-consistency measure assuming that corresponding pixels have the same color. Besides this restrictive brightness constancy assumption the method works only for camera configurations where all scene points lie outside the convex hull of the camera centers. This configuration constraint was relaxed in a generalization of the voxel coloring by Kutulakos and Seitz [2000]. They introduce the *photo hull* which encloses the set of all photo-consistent shapes and compute it by *space carving* which means they prune away empty voxels from the volume.

2.1.2 Surface from Images

A popular approach in surface reconstruction from sample points is surface evolution which starting from an initial surface S_0 aims to find the surface S that minimizes an energy $E(S)$. Hernandez *et al*. [2004] apply this concept to surface reconstruction from

images. They use silhouette information in the images to compute a 3D convex hull. Then they apply an octree-based carving method followed by a marching tetrahedron meshing algorithm and a mesh simplification to find the initial surface. The external forces that drive the surface evolution fuse texture and silhouette information. The internal force implements a regularization on the surface.

Gargallo *et al*. [2005] use a Bayesian approach that also leads to an energy minimization problem. They introduce a visibility prior and a multiple depth map prior that takes into account the different view points and aims for generally smooth depth maps still allowing for sharp discontinuities. The posterior probability is then maximized with the generalized Expectation Maximization algorithm where the maximum is approached by gradient descent. Later on, they compute the exact gradient of the reprojection error function and use this error function for gradient descent surface evolution [Gargallo et al. 2007].

Graph cut [Boykov et al. 2001] based methods aim to find the surface with minimal energy as well. After the development of fast energy minimization algorithms based on graph cuts, this concept was first successfully applied to the original two-view stereo. Kolmogorov and Zabih generalized it to multi-camera scene reconstruction [2002]. They formulate an energy minimization consisting of a data term that imposes photo-consistency, a smoothness term, and a visibility term. The resulting energy minimization problem is NP-hard to minimize exactly but with a graph cut an approximate solution can be computed. A year later, Boykov and Kolmogorov [2003] showed how a minimum surface under an arbitrary Riemannian metric can be found using graph cut algorithms. This work inspired numerous graph cut based multi-view stereo algorithms including [Hornung and Kobbelt 2006a, Hornung and Kobbelt 2006b, Lempitsky and Boykov 2007, Sinha et al. 2007, Sormann et al. 2007, Vogiatzis et al. 2007].

Pons *et al*. [2007] propose a variational model for multi-view stereo reconstruction from video sequences using a global image-based matching score. They simultaneously estimate the shape of the object and the 3D scene flow solving the image registration task. The method inherently requires a small baseline between the images which makes it unsuitable for standard multi-view stereo data sets.

For urban outdoor and indoor scenes piecewise planar models have become popular as well. The common intuition is that man-made scenes mainly consist of piecewise planar surfaces, often perpendicular to each other. Existing methods either fit planes to the reconstructed multi-view stereo points [Furukawa et al. 2009] or directly estimate piecewise planar depth maps using a Markov Random Field optimization [Sinha et al. 2009]. Gallup *et al*. [2010] take a similar approach but additionally segment the images

into planar and non-planar regions. From an initial set of depth maps they create plane hypotheses using a RANSAC method and use graph cut labeling to assign each pixel to a plane that is consistent between different views. Hereby they allow for a *non-plane* label preserving the original depth values.

2.1.3 Sample Points from Images

Several multi-view stereo methods compute 3D points, either as depth maps with connectivity information or as an unordered point cloud. If a triangle mesh is desired they often apply a standard surface reconstruction algorithm and discard the potentially existing connectivity information. Goesele *et al*. [2007] apply multi-view stereo on images from community photo collections using a two-stage view selection. A surface growing approach with varying disc size dependent on image texture is proposed by Habbecke and Kobbelt [2007]. Bradley *et al*. [2008] start with scaled window matching and apply a filtering to obtain a noise-reduced point cloud. The widely used technique by Furukawa and Ponce [2010] reconstructs a dense set of patches that represents the surface. A clustering approach allows for applicability on extremely large photo collections [Furukawa et al. 2010]. They create overlapping clusters, process them in parallel, and finally merge the individual reconstructions.

Labatut *et al*. [2007] first create a quasi-dense feature point cloud. Then they apply a 3D Delaunay triangulation and extract the final surface as a subset of faces that minimizes an energy taking into account visibility, photo-consistency and smoothness. The minimum is found using a graph cut. This method has been extended using a different energy term [Labatut et al. 2009] and to work on high-resolution images and large-scale scenes adding Difference of Gaussians (DoG) features and Harris points [Harris and Stephens 1988] to obtain a denser point cloud [Hiep et al. 2009]. Bailer *et al*. [2012] use the same Delaunay based optimization [Labatut et al. 2009] but create the point cloud differently. They first compute depth maps, filter them to remove erroneous points, and then project the points into 3D space, and improve the point cloud using a moving-least squares variant. Jancosek *et al*. [2009] create filtered meshes from grown patches and detect overlapping areas. Their final representation is not a closed surface but is composed of locally consistent meshes with minimum overlap.

2.2 Surface Reconstruction from Sample Points

In the beginning, sample points were the result of a range scanning process from a single or multiple viewpoints. In recent times, however, sample points also originate from

multi-view stereo methods, either directly in 3D space or as transformed depth maps. Many methods do, however, assume that 3D sample points with only their positions are given. Some assume that normal information (oriented points) or the direction to the sensor is additionally given. Finally, there are methods specifically tailored to a particular application setting such as merging depth maps that originate from multi-view stereo matching between images. Since we think the underlying concept is a vital difference between the methods we decided to loosely sort the related work by this criterium.

There is also a brand-new benchmark by Berger *et al*. [2013] that compares popular methods on several data sets using different error metrics. The generation of the point clouds they provide is designed to mimic laser scans, so the points are more or less regularly sampled and single-scale.

2.2.1 Delaunay-based Methods

The idea of using a Delaunay triangulation for surface reconstruction was introduced by Boissonnat [1984]. The *Delaunay triangulation* subdivides the convex hull of the sample points and is unique under certain sampling conditions. The dual of the Delaunay triangulation is the *Voronoi diagram* which subdivides the space into convex cells. Each cell can be associated with exactly one sample point.

Among the various Delaunay-based methods the most popular are perhaps the *Crust* [Amenta and Bern 1999, Amenta et al. 2001] and its successor the *Cocone* [Amenta et al. 2002, Dey and Goswami 2003]. Both exploit the structure of the Voronoi diagram of the input points to remove triangles that do not belong to the surface. These methods work well for densely sampled point clouds but fail if sampling is sparse. A greedy method was presented by Cohen-Steiner and Da [2004]. Starting from a seed triangle they grow a surface by adding always the most plausible Delaunay triangles under the assumption that normals vary smoothly over the surface. In this way they prevent topological singularities and can even handle non-closed surfaces with boundaries. Dey *et al*. [2009] present an algorithm that guarantees an isotopic reconstruction of surfaces with boundaries if the sampling is noise-free. Alliez *et al*. [2007] combine a Delaunay-based approach with an implicit surface representation using a spectral method. Labatut *et al*. [2009] define an energy that consists of a visibility term, taking the direction to the sensor into account, and a surface quality cost. The energy can be interpreted as costs of removing edges in a graph, that correspond to faces of the Delaunay triangulation, and a minimum cut yields the reconstructed surface.

The main advantage of Delaunay- or Voronoi-based reconstruction techniques is that

they allow for a theoretical analysis proving the reconstruction quality, *e.g.*, guaranteed geometric features. It is, however, only suited if the sampling is noise-free and dense enough. More details about Delaunay-based surface reconstruction can be found in the survey by Cazals and Giesen [2006].

2.2.2 Surface Evolution

Level set-based surface reconstruction uses deformable models. Starting from an initial shape they iteratively alter the shape to minimize an energy. One can separate the methods into *ballooning* techniques that grow the surface from the inside [Cohen and Cohen 1993, Zhao et al. 2001, Sharf et al. 2006], and *shrinking* techniques growing from the outside [Esteve et al. 2005]. Tagliasacchi *et al*. [2011] set up a surface evolution framework based on a level set formulation that incorporates weak volumetric priors in order to better reconstruct objects with many concavities. Level set-based formulations are also used for surface reconstruction from range images. For example, Whitaker [1998] uses a statistical formulation of the 3D reconstruction problem and represents the surface as the level set of a discretely sampled scalar function. This function is altered, which mimics deforming the surface, in order to maximize a posterior probability including a noise model and a surface prior.

2.2.3 Implicit Surface Representation

Many methods compute an implicit surface representation where the zero level set represents the unknown surface. This can be extracted using marchings cubes [Lorensen and Cline 1987] or other contouring algorithms [Schaefer et al. 2007, Manson and Schaefer 2010]. An implicit representation of the surface is given by the signed distance field. Hoppe *et al*. [1992] approximate the signed distance for a point by computing the distance to the least squares plane of its k-nearest neighbors. Carr *et al*. [2001] use polyharmonic radial basis functions (RBFs) as implicit surface representation. They also construct a signed distance function but subsequently fit a radial basis function to the distance field. Ohtake *et al*. [2003b] introduce a hierarchical reconstruction approach where they use globally and locally supported radial basis functions to implicitly represent the surface. On the given point cloud they first apply a spatial downsampling to construct a coarse-to-fine point set hierarchy. They then successively interpolate the sets starting from the coarsest level. For each finer level they only interpolate the offset of the interpolating function computed at the previous level. Another implicit surface representation is the characteristic function of the object defined, *e.g.*, as being 1 out-

side and -1 inside the object. Kazhdan [2005] computes the Fourier coefficients of the characteristic function. In a follow-up paper he and his colleagues turn the problem of recovering the characteristic function into a spatial Poisson problem [Kazhdan et al. 2006]. *Poisson surface reconstruction* is still widely used and serves as a reference not least because of the publicly available implementation. Bolitho *et al*. [2007] present an out-of-core solution for huge Poisson systems based on a multi-level streaming representation that increases reconstruction speed for large 3D scans. Very recently, Kazhdan and Hoppe [2013] extended the original algorithm by adding an interpolation constraint. This leads to surfaces that better follow the input data and thus better model sharp details. Very similar to [Kazhdan 2005] Manson *et al*. [2008] model a smoothed version of the characteristic function using wavelets. This is much faster because wavelets typically have local support in contrast to the Fourier basis functions. Additionally, the hierarchical structure of wavelets can be exploited in a streaming surface reconstruction implementation. Calakli and Taubin [2011] provide a generalized framework where they represent the signed distance field using any linearly parameterized family of smooth basis functions. They turn the surface reconstruction task into an energy minimization problem and show how this can be transformed into a linear system of equations. Taylor [2003] effectively computes the characteristic function of the object but in a different way. The underlying idea is to infer information about free space from each sample point and afterwards triangulate its boundary.

Dong *et al*. [2011] first define a general variational model for surface reconstruction similar to models used for image restoration. The final surface is hereby represented using an unsigned distance field. They then propose a wavelet frame-based model that can be interpreted as a certain discretization to the variational model. The projection-based moving least squares technique [Levin 2004] defines the surface as the invariant of a parametric fit procedure. Alexa *et al*. [2003] introduced this concept to computer graphics computing the point-based representation of the moving least squares surface for rendering purposes. In the meantime several variants of this technique have been proposed [Shen et al. 2004, Fleishman et al. 2005]. Based on mean curvature motion Digne *et al*. [2011] define a smoothing operator on raw point clouds. Successive applications on the input points lead to a scale space representation of the surface. They triangulate the coarsest scale using a standard meshing algorithm and transport the vertices back to their original positions. The multi-level partition of unity [Ohtake et al. 2003a] is a local implicit surface representation. Piecewise quadratic functions that describe the local surface shape are blended together using weighting functions resulting in an approximation of the true signed distance function. Nagai *et al*. [2009] define

gradient operators on partition of unity implicits and apply Laplacian smoothing on the gradient field in order to cope with noisy data.

The surface reconstruction algorithm *VRIP* [Curless and Levoy 1996] was the first to incorporate visibility information using space carving. Like many methods Curless and Levoy reconstruct the signed distance field. However, for each point they take the direction to the sensor into account to better model the positional uncertainty in the acquisition process. In an energy minimization framework Zach *et al*. [2007] incorporate a total variation regularization on the distance field and use the L_1—norm for data fidelity to gain robustness against outliers. The input to their method are truncated distance fields generated from the depth maps similar to Curless and Levoy [1996]. Taking the visibility information into account is also the key concept in the cone carving method [Shalom et al. 2010]. They compute an improved signed distance by associating each point with an estimated visibility cone that carves outside space of the object. Fuhrmann and Goesele [2011] introduce a hierarchical signed distance field on a voxel grid where the hierarchies map different surface scales. Starting from triangulated depth maps they construct the hierarchical signed distance field similar to *VRIP* but each triangle only affects a certain hierarchy level depending on an estimated scale. During a regularization step coarse scale information is discarded when reliable fine scale information is available. To extract the isosurface they apply a Delaunay triangulation of the adaptive voxel grid and use a variant of Marching Tetrahedra [Doi and Koide 1991].

2.2.4 Other Methods

In the early *mesh zippering* approach [Turk and Levoy 1994] range scans are triangulated, redundant triangles removed, and the meshes pairwise *zippered* together at the boundaries. Finally, vertex positions are refined according to the original range scans.

Scattered data reconstruction is also continuously researched in the field of approximation theory. The objective here is mostly to reconstruct one- or two-dimensional functions which corresponds to height fields. A common approach is the approximation in B-spline and wavelet spaces [Pastor and Rodríguez 1999, Johnson et al. 2009]. Recently, Ji *et al*. [2010] proposed a method where they use tight wavelet frames to reconstruct the surface given range data of a single view. This allows for the reconstruction of sharp edges and increases robustness to noise and outliers. These and other methods in scattered data interpolation are related to one of our proposed surface reconstruction algorithms. In contrast to our work, however, they do not tackle the problem of multi-scale input data.

Using a Bayesian approach Jenke *et al*. [2006] reconstruct a noise-free and well sam-

pled point cloud that is most likely to be a subsampled version of the true surface. For meshing they use a variant of moving least squares combined with a standard implicit surface reconstruction method [Hoppe et al. 1992]. Gal *et al*. [2007] take the concept of recovering a clean point cloud one step further and incorporate local shape priors from a data base of example shapes. The main drawback of these Bayesian approaches is that they are computationally very expensive and thus not applicable on large data sets.

3 Foundations

Contents

THIS chapter gives an overview of the general surface reconstruction pipeline used in this thesis (similar to Goesele *et al*. [2007], see Figure 3.1). The input is an unordered set of images. The photos can be taken with a consumer camera or even downloaded from the Internet, *e.g*., by choosing images on photo sharing sites such as flickr that match a particular tag. The images are *registered* first to recover the internal and external camera parameters and to create a sparse point cloud representing the scene. In the next step, we compute a depth map for each registered image where we try to assign a depth value for each pixel. Finally, the depth maps are fused to a global consistent model such as a triangle mesh. Note that some multi-view stereo methods directly reconstruct a global model instead of computing a local representation first. The following sections elaborate on the separate steps, provide details about the specific algorithms used in this thesis, and describe how the contributions fit into the pipeline.

3.1 Structure-from-Motion

In the first step the images are registered, *i.e*., a spatial relationship between the images is established. We use the structure-from-motion technique introduced by Snavely *et al*. [2006, 2008]. They detect keypoints (*e.g*., using SIFT [Lowe 2004]) in all images and match the keypoint descriptors in order to connect two images. In a RANSAC process they estimate a fundamental matrix on the basis of eight matching key points. The matches are then organized into tracks to connect multiple images. Starting from two very well matching images, more images along the tracks are added to build a scene. During this process a *bundle adjustment* [Triggs et al. 2000] optimizes for consistent

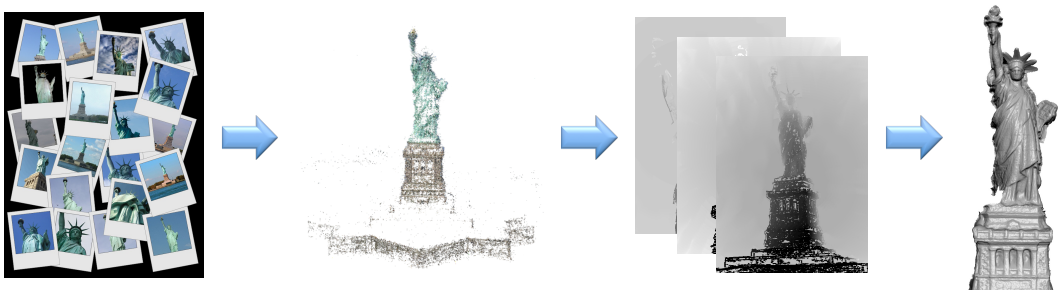


Figure 3.1: Reconstructing a closed surface from images: Starting from a set of images (*left*) a sparse scene representation together with the camera parameters is recovered using a *structure-from-motion* technique. Then for each view a dense depth map is computed using *multi-view stereo*. Finally, a dense triangle mesh is reconstructed by fusing the depth maps (*right*).

camera parameters. The triangulated feature points already provide a sparse scene representation (see Figure 3.1 *(middle left)*).

The computed 3D feature points and their projections in the images are input to the multi-view stereo in the next step. The computed camera positions are naturally not perfect but most of the time good enough and rarely contain outliers. We therefore take the information as is and did not try to model or even correct the potential errors.

3.2 Multi-View Stereo

Starting from the 3D feature point cloud created in the previous step the multi-view stereo algorithm [Goesele et al. 2007] determines a depth value for each reconstructable pixel in every image of the scene. For a given reference image the algorithm first determines neighbor images in a *global view selection* that are suited for reconstruction. In a region growing fashion a photo-consistency minimization recovers optimal depth and normal of a small 3D patch centered around the corresponding 3D point of the current pixel. If the optimization terminates and a threshold in photo-consistency with four neighboring views (chosen from the *local view selection*) is met, the values are accepted. Otherwise the depth of that pixel is declared as unknown (see Figure 3.1 *(middle right)*).

In contrast to structure-from-motion the reconstructed points from multi-view stereo contain a lot of noise and outliers. On the other hand, these points have the capability to capture fine surface details and provide a considerably denser scene representation. In Chapter 5, we propose a theoretical model for the multi-view stereo reconstruction process that allows us to predict the systematic error concerning fine scale details. The subsequent chapter proposes a weighted photo-consistency optimization in order to achieve

a better frequency behavior.

3.3 Surface Reconstruction

In the previous step a set of depth maps has been computed which provide local scene representations that are not necessarily consistent with each other. The final step is to compute a global and consistent scene representation (see Figure 3.1 *(right)*). Popular methods that proved suitable for this task and with source code available online are VRIP [Curless and Levoy 1996], Poisson surface reconstruction [Kazhdan et al. 2006], and the recent Depth Map Fusion algorithm [Fuhrmann and Goesele 2011].

This thesis introduces two distinct approaches that both take the combined set of 3D points from all depth maps as input. The first method presented in Chapter 4 estimates the size of the 3D patch used during photo-consistency optimization. This serves as a measure of confidence of that sample and allows for fine scale samples to steer the reconstruction of fine details. The second method proposed in Chapter 7 assumes weighted photo-consistency optimization and takes as additional input the (estimated) weighting function.

4 Hierarchical Surface Reconstruction

Contents

ROBUST surface reconstruction from sample points is a challenging problem, especially for real-world input data. We present a new hierarchical surface reconstruction based on volumetric graph cuts that incorporates significant improvements over existing methods. One key aspect of our method is, that we exploit the footprint information which is inherent to each sample point and describes the underlying surface region represented by that sample. We interpret each sample as a vote for a region in space where the size of the region depends on the footprint size. In our method, sample points with large footprints do not destroy the fine detail captured by sample points with small footprints. The footprints also steer the inhomogeneous volumetric resolution used locally in order to capture fine detail even in large-scale scenes. Similar to other methods our algorithm initially creates a crust around the unknown surface. We propose a crust computation capable of handling data from objects that were only partially sampled, a common case for data generated by multi-view stereo algorithms. Finally, we show the effectiveness of our method on challenging outdoor data sets with samples spanning orders of magnitude in scale.

Figure 4.1: *Left:* An input image to multi-view stereo reconstruction. *Middle:* The reconstructed depth map visualized in gray values (white: far, black: near). *Right:* The triangulated depth map rendered from a slightly different view point.

4.1 Introduction

Reconstructing a surface mesh from sample points is a problem that occurs in many applications, including surface reconstruction from images as well as scene capture with triangulation or time-of-flight scanners. Our work is motivated by the growing capabilities of multi-view stereo (MVS) techniques [Seitz et al. 2006, Goesele et al. 2007, Habbecke and Kobbelt 2007, Furukawa et al. 2010] that achieve remarkable results on various data sets.

Traditionally, surface reconstruction techniques are designed for fairly high-quality input data. Measured sample points, in particular samples generated by MVS algorithms, are, however, *noisy* and contain *outliers*. Figure 4.1 shows an example reconstructed depth map that we use as input data in our method. Furthermore, sample points are often non-uniformly distributed over the surface and entire regions might not be represented at all. Recently, Hornung and Kobbelt presented a robust method well suited for noisy data [2006b]. This method generates optimal low-genus watertight surfaces within a crust around the object using a volumetric graph cut. Still, their algorithm has some major limitations regarding crust generation, sample footprint, and missing multi-resolution reconstruction which we address in this chapter.

Hornung and Kobbelt create a surface confidence function based on unsigned distance values extracted from the sample points. The final surface S is obtained by optimizing for maximum confidence and minimal surface area. As in many surface reconstruction algorithms, the footprint of a sample point is completely ignored when computing the confidence. Every sample point, regardless of how it was obtained, inherently has a *footprint*, the underlying surface area taken into account during the measurement (see Figure 4.2). The size of the footprint indicates the sample point's capability to capture surface details. A method that outputs sample points with different footprints was proposed by Habbecke and Kobbelt [2007]. They represent the surface with surfels (surface

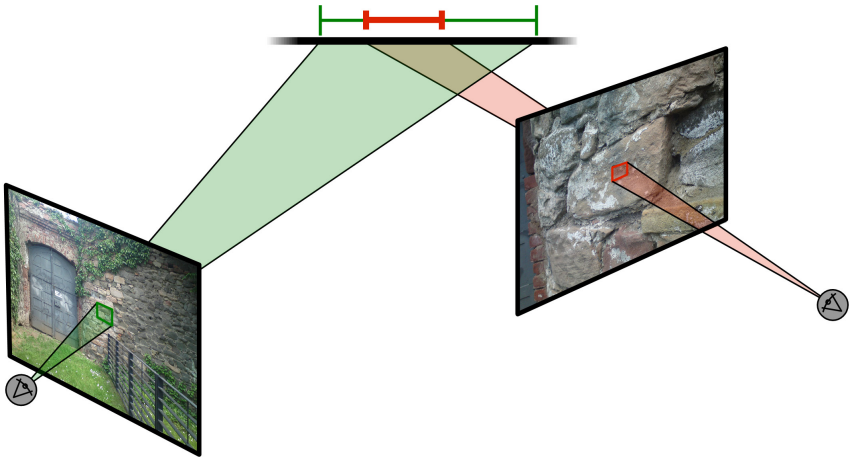


Figure 4.2: Visualization of the *footprint* of a sample point: A certain pixel in the left image covers a significantly larger area than a corresponding pixel in the right image.

elements) of varying size depending on the image texture. Furukawa *et al*. [2010] consider footprints to estimate reconstruction accuracy and Fuhrmann and Goesele [2011] build a hierarchical signed distance field where they insert samples on different scales depending on their footprint. However, both methods effectively discard samples with large footprints prior to final surface extraction. In this chapter, we propose a different way to model the sample footprint during the reconstruction process. In particular, we create a modified confidence map where samples contribute differently depending on their footprints.

The confidence map is only evaluated inside a *crust*, a volumetric region around the sample points. In [Hornung and Kobbelt 2006b], the crust computation implicitly segments the boundary of the crust into *interior* and *exterior*. The final surface separates interior from exterior. This crust computation basically works only for completely sampled objects. Even with their proposed workaround (estimating the medial axis), the resulting crust is still not applicable to many data sets. Such a case is illustrated in Figure 4.3, where no proper interior component can be computed. This severely restricts the applicability of the entire algorithm. We propose a different crust computation that separates the crust generation from the crust segmentation process, extending the applicability to a very general class of input data.

Finally, as Hiep *et al*. [2009] pointed out, volumetric methods such as [Hornung and Kobbelt 2006b] relying on regular volume decomposition are not able to handle large-scale scenes. To overcome this problem our algorithm reconstructs on a locally adaptive volumetric resolution and finally extracts a watertight surface. This allows us to

reconstruct fine details even in large-scale scenes such as the Citywall data set (see Figure 4.11).

The remainder of this chapter is organized as follows: First, we review previous work (Section 4.2) and give an overview of our reconstruction pipeline (Section 4.3). Details of the individual steps are explained in Sections 4.4–4.7. Finally, we present results of our method on standard benchmark data as well as challenging outdoor scenes (Section 4.8) and wrap up with a conclusion and an outlook on future work (Section 4.9).

4.2 Related Work

Surface reconstruction from (unorganized) points

Surface reconstruction from unorganized points is a large and active research area. One of the earliest methods was proposed by Hoppe *et al*. [1992]. Given a set of sample points, they estimate local tangent planes and create a signed distance field. The zero-level set of this signed distance field, which is guaranteed to be a manifold, is extracted using a variant of the marching cubes algorithm [Lorensen and Cline 1987].

If the sample points originate from multiple range scans, additional information is available. VRIP [Curless and Levoy 1996] uses the connectivity between neighboring samples as well as the direction to the sensor when creating the signed distance field. Additionally, it employs a cumulative weighted signed distance function allowing it to incrementally add more data. The final surface is again the zero-level set of the signed distance field. A general problem of signed distance fields is that local inconsistencies of the data lead to surfaces with undesirably high genus and topological artifacts. Zach *et al*. [2007] mitigate this effect. They first create a signed distance field for each range image and then compute a regularized field u approximating all input fields while minimizing the total variation of u. The final surface is the zero-level set of u. Their results are of good quality, but the resolution of both, the volume and the input images, is very limited. In their very recent paper, Fuhrmann and Goesele [2011] introduce a depth map fusion algorithm that takes sample footprints into account. They merge triangulated depth maps into a hierarchical signed distance field similar to VRIP. After a regularization step, basically pruning low-resolution data where reliable higher-resolution data is available, the final surface is extracted using marching tetrahedra. Our method does not rely on triangulated depth maps and tries to merge all data samples while never discarding information from low-resolution samples. Another recent work taking unorganized points as input is called cone carving and is presented by Shalom *et al*. [2010]. They associate each point with a cone around the estimated normal to carve free space

and obtain a better approximation of the signed distance field. This method is in a way characteristic for many surface reconstruction algorithms in the sense that it is designed to work on raw scans from a commercial 3D laser scanner with rather good quality. Such methods are often not able to deal with the lower quality data generated by MVS methods from outdoor scenes containing a significant amount of noise and outliers.

Kazhdan *et al*. [2006] reformulate the surface reconstruction problem as a standard Poisson problem. They reconstruct an indicator function marking regions inside and outside the object. Oriented points are interpreted as samples of the gradient of the indicator function, requiring accurate normals at each sample point's position which are usually not present in MVS data. The divergence of the smoothed vector field, represented by these oriented points, equals the Laplacian of the indicator function. The final surface is extracted as an iso-surface of the indicator function using a variant of the marching cubes algorithm. Along these lines, Alliez *et al*. [2007] use the normals to derive a tensor field and compute an implicit function whose gradients best approximate that tensor field. Additionally, they present a technique, called Voronoi-PCA, to estimate unoriented normals using the Voronoi diagram of the point set.

Graph cut based surface reconstruction

Boykov and Kolmogorov [2003] introduce the idea of reconstructing surfaces by computing a cut on a graph embedded in continuous space. They also show how to build a graph and set the edge weights such that the resulting surface is minimal for any anisotropic Riemannian metric. Hornung and Kobbelt [2006a] use the volumetric graph cut to reconstruct a surface given a photo-consistency measure defined at each point of a predefined volume space. They propose to embed an octahedral graph structure into the volume and show how to extract a mesh from the set of cut edges. In a follow-up paper [Hornung and Kobbelt 2006b], they present a way to compute confidence values from a non-uniformly sampled point cloud and improve the mesh extraction procedure.

An example of using graph cuts in multi-view stereo is the work of Sinha *et al*. [2007]. They build an adaptive multi-resolution tetrahedral mesh where an estimated photo-consistency guides the subdivision. The final graph cut is performed on the dual of the tetrahedral mesh followed by a photo-consistency driven mesh refinement. Labatut *et al*. [2009] build a tetrahedral mesh around points merged from multiple range images. They introduce a surface quality term and a surface visibility term that takes the direction to the sensor into account. From an optimal cut, which minimizes the sum of the two terms, a labeling of each tetrahedra as inside or outside can be inferred. The final mesh consists of the set of triangles separating the tetrahedra according to their labels.

Hiep *et al*. [2009] replace the point cloud obtained from multiple range images with a set of 3D features extracted from the images. The mesh obtained from the tetrahedral graph cut is refined mixing photo-consistency in the images and a regularization force. However, none of the existing graph cut based surface reconstruction algorithms properly incorporates the footprint of a sample.

4.3 Overview

The input of our algorithm is a set of *surface samples* representing the scene (Figure 4.3a). Each surface sample consists of its position, footprint size, a scene surface normal approximation, and an optional confidence value. A cubic bounding box is computed from the input points or given by the user.

First, we determine the *crust*, a subset of the bounding volume containing the unknown surface. All subsequent computations will be performed inside this crust only. Furthermore, the boundary of the crust is partitioned into *interior* and *exterior*, defining interior and exterior of the scene (Figure 4.3b). Inside the crust we compute a *global confidence map*, such that points with high confidence values are likely to lie on the unknown surface. Each sample point adds confidence to a certain region of the volume. The size of the region and the confidence peak depend on the sample point's footprint size. Effectively, every sample point adds the same total amount of confidence to the volume but spread out differently. A volumetric graph is embedded inside the crust where graph nodes correspond to voxels and graph edges map the 26-neighborhood. A minimal cut on this graph separates the voxels into interior and exterior representing the optimal surface at this voxel resolution (Figure 4.3c). The edge weights of the graph are chosen such that the final surface minimizes surface area while maximizing confidence.

We then identify surface regions with sampled details too fine to be adequately represented on the current resolution. Only these regions are subdivided, the global confidence map is resampled, and the graph cut is computed on a higher resolution (Figure 4.3d+e). We repeat this process iteratively until eventually all fine details were captured. Finally, we extract the surface in the irregular voxel grid using a combination of marching cubes and marching tetrahedra. This results in a multi-resolution surface representation of the scene, the output of our algorithm (Figure 4.3f).

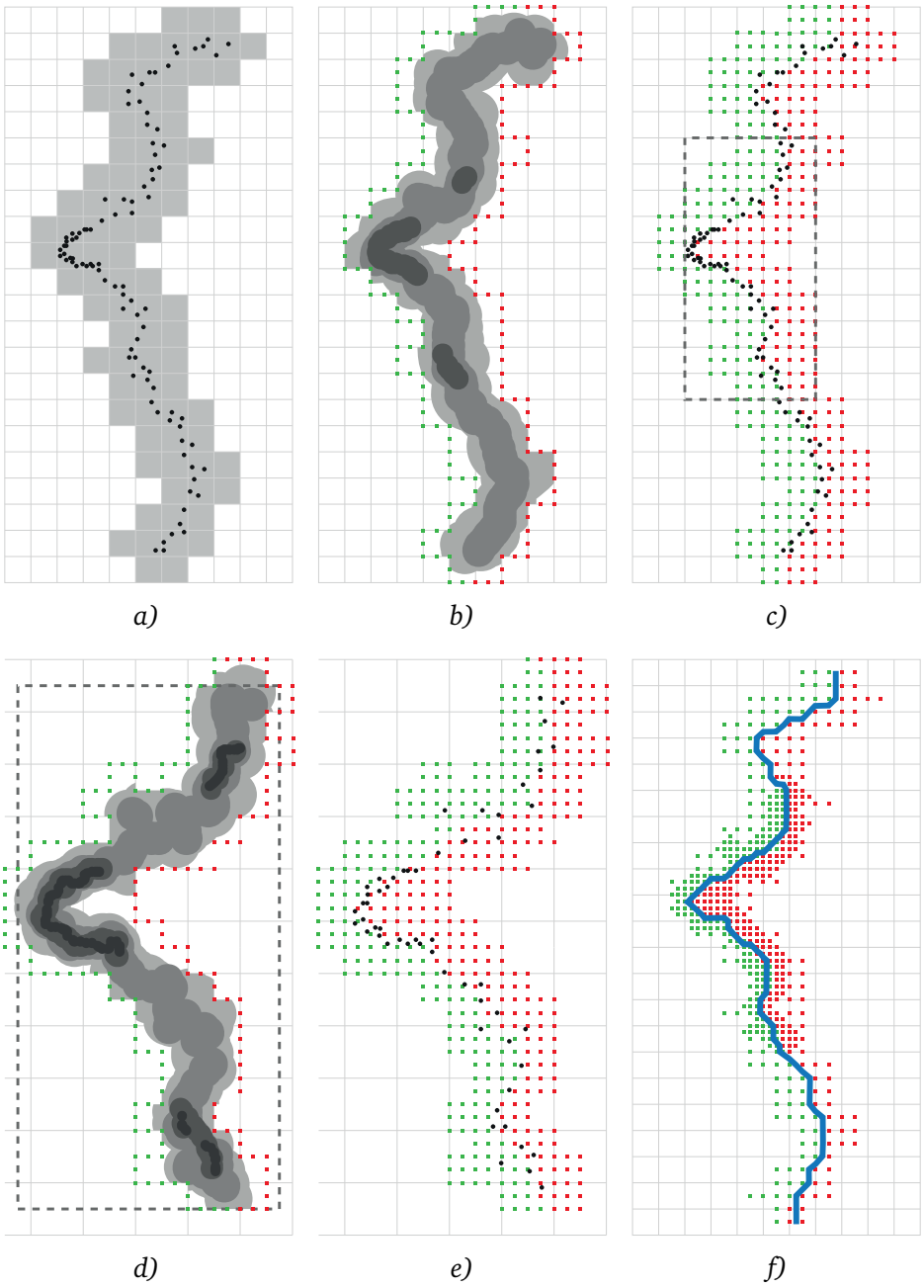


Figure 4.3: Overview of our reconstruction pipeline. *a)* We compute a crust around the input samples of different footprints and varying sampling density. *b)* We segment the crust into *interior* (red) and *exterior* (green) and compute the global confidence map (GCM) to which each input sample contributes. *c)* A minimal cut on the embedded graph segments the voxel corners representing the surface with maximum confidence while minimizing surface area. We mark the areas with high-resolution samples (dashed black box) and iteratively increase resolution therein. *d+e)* In the increased resolution area we re-evaluate the GCM and perform the graph cut optimization. *f)* Finally, an adaptive triangle mesh is extracted from the multi-resolution voxel corner labeling.

4.4 Crust Computation

We subdivide the cubic bounding box into a regular voxel grid. For memory efficiency and to easily increase the voxel resolution, this voxel grid is represented by an octree data structure. Our algorithm iteratively treats increasing octree levels (finer resolution) starting with a user-defined low octree level ℓ_0, i.e., with a coarse resolution.

The crust $V_{crust} \subset V$ is a subset of voxels that contains the unknown surface. The crust computation is an important step in the algorithm for several reasons: The shape of the crust constrains the shape of the reconstructed surface. Furthermore, the crust has to be sufficiently large to contain the optimal surface and on the other hand as narrow as possible to reduce computation time and memory cost. We split the crust computation into two parts. First, the crust is generated, then the boundary of this crust is segmented to define interior and exterior of the scene (see Figure 4.4 for an overview).

Crust Generation We initialize the crust on level ℓ_0 with the set of voxels on the parent octree level $\ell_0 - 1$ containing surface samples. We dilate this sparse set of voxels several times over the 6-neighborhood of voxels, followed by a morphological closing operation (Figure 4.4a). The number of dilation steps is currently set by the user, but the resulting crust shape can be immediately inspected, as the crust generation is fast on the low initial resolution. Subsequently, these voxels $v \in V_{crust}^{\ell_0-1}$ are once regularly subdivided to obtain the initial crust $V_{crust}^{\ell_0}$ for further computations on level ℓ_0.

Crust Segmentation In this step our goal is to assign labels *interior* and *exterior* to all boundary voxel corners on level ℓ_0 to define the interior and exterior of the scene. In the following, we define $\partial V_{crust}^{\ell}$ to be the set of boundary voxels on level ℓ. We start by determining labels for voxel corners v_f that lie on the midpoints of boundary faces of parent crust voxels $v \in \partial V_{crust}^{\ell_0-1}$. The labels are determined by comparing a surface normal estimate $\vec{n}_v^{surf}$ for parent voxel v with the normals of the boundary faces $\vec{n}_{v_f}^{crust}$. The surface normal is computed for each crust voxel by averaging the normals of all sample points inside the crust voxel. Crust voxels that do not contain surface samples obtain their normal estimate through propagation during crust dilatation (Figure 4.4b). We determine the initial labels on the crust boundary by

$$label(v_f) = \begin{cases} exterior, & \text{if} \quad \vec{n}_{v_f}^{crust} \cdot \vec{n}_v^{surf} \geq \tau \\ interior, & \text{if} \quad \vec{n}_{v_f}^{crust} \cdot \vec{n}_v^{surf} \leq -\tau \\ \text{unknown}, & \text{otherwise} \end{cases} \tag{4.1}$$

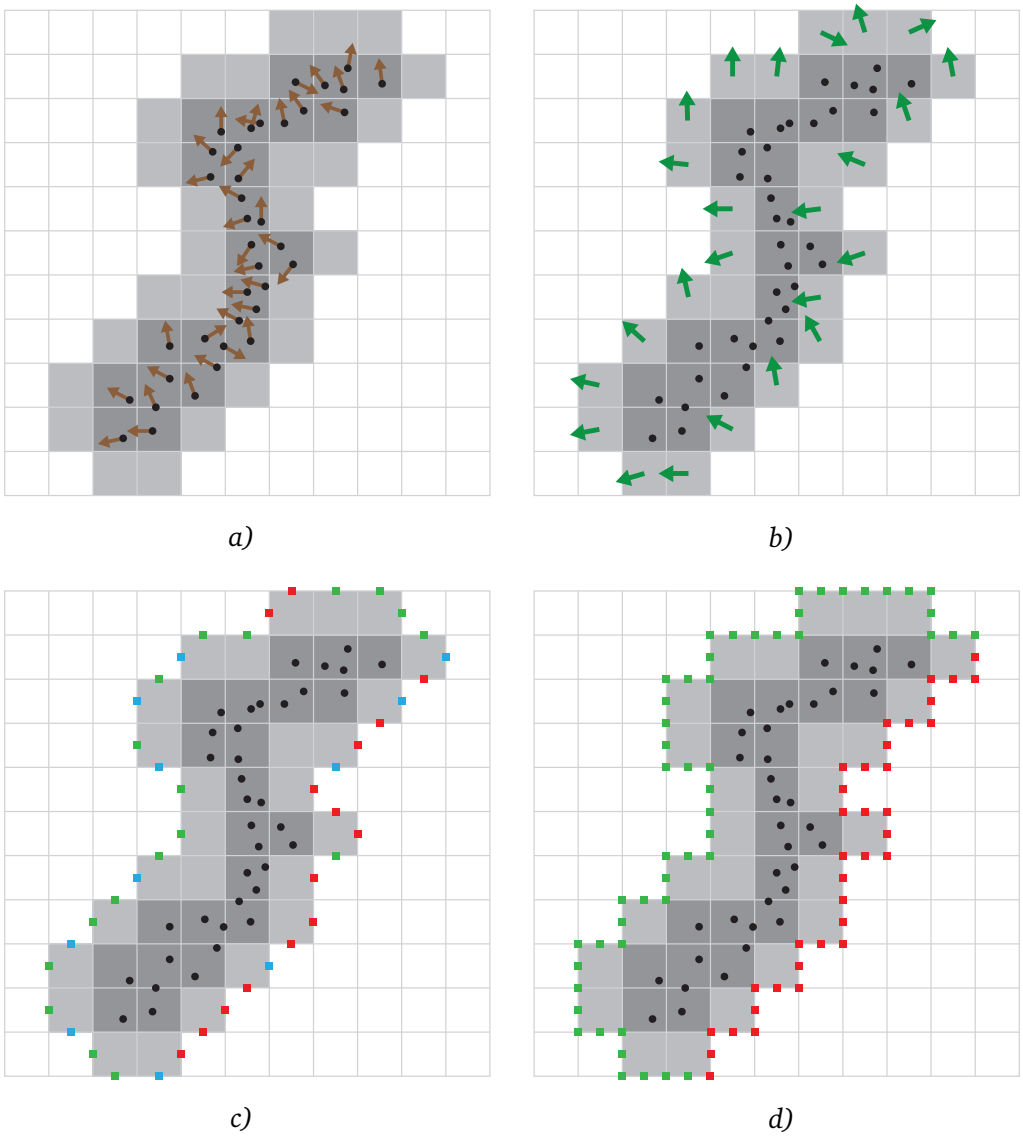


Figure 4.4: Initial crust computation for lowest resolution: *a)* We initialize the crust with voxels containing sample points and dilate several times. *b)* Surface normals are computed for each voxel. *c)* The comparison of surface normals with the face normals of the crust voxels defines an initial labeling into *interior* (red), *exterior* (green), and *unknown* (blue). *d)* An optimization yields a homogenous crust surface segmentation.

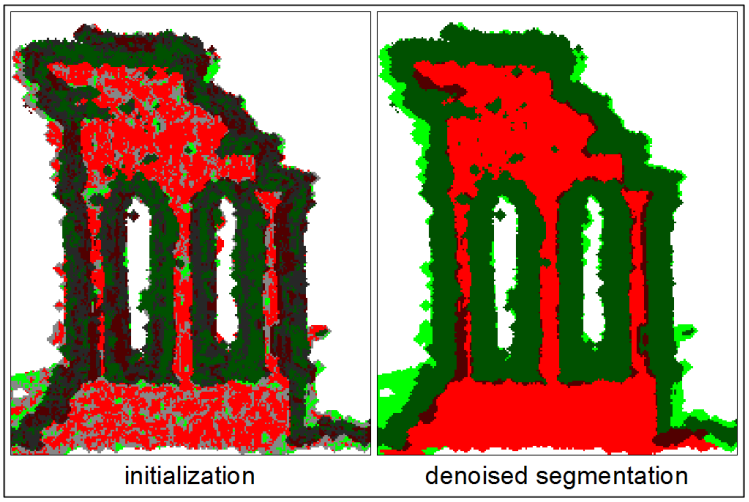


Figure 4.5: Visualization of the crust surface for the Temple (cut off perpendicular to the viewing direction). The color is similar to Figure 4.4. Light shaded surfaces are seen from the front, dark shaded ones are seen from the back.

with $\tau \in (0,1)$ (Figure 4.4c). We used $\tau = 0.75$ in all experiments.

By now we have just labeled a subset of all voxel corners on level ℓ_0 (Figure 4.4c). Furthermore, since surface normal information of the samples may only be a crude approximation, this initial labeling is noisy and has to be regularized. We cast the problem of obtaining a homogenous labeling of the crust surface into a 2D binary image denoising problem solved using graph cut optimization as described by Boykov and Veksler [Boykov and Veksler 2006]. We build a graph with a node per voxel corner in $\partial V_{crust}^{\ell_0}$ and a graph edge connecting two nodes if the corresponding voxel corners share a voxel edge. Additionally, 'diagonal' edges are inserted that connect the initially labeled corners in the middle of parent voxel faces with the four parent voxel corners. We also add two terminal nodes *source* and *sink* together with further graph edges connecting each node to these terminals. Note that this graph is used for the segmentation of the crust on the lowest resolution level ℓ_0 only and should not be confused with the graphs used for surface reconstruction on the different resolutions.

All edges connecting two non-terminal nodes receive the same edge weight w. Edges connecting a node n with a terminal node receive a weight depending on the labeling of the corresponding voxel corner v_c, where unlabeled voxel corners are treated as

unknown:

$$w_n^{source} = \begin{cases} \mu & \text{if } v_c \text{ is labeled } interior \\ 1-\mu & \text{if } v_c \text{ is labeled } exterior \\ \frac{1}{2} & \text{if } v_c \text{ is unknown} \end{cases} \tag{4.2}$$

$$w_n^{sink} = 1 - w_n^{source} \tag{4.3}$$

for a constant $\mu \in (0, \frac{1}{2})$. With these edge weights the *exterior* is associated with *source*, *interior* with *sink*. A cut on this graph assigns each node either to the *source* or to the *sink* component and therefore yields a homogeneous segmentation of the boundary voxel corners of $\partial V_{crust}^{\ell_0}$ (Figure 4.4d and Figure 4.5 right). We used $w = 0.5$ and $\mu = 0.25$ in all experiments.

If two neighboring crust voxel corners obtained different labels, the reconstructed surface is forced to pass between them, as it has to separate *interior* from *exterior*. The denoising minimizes the number of such occurrences and therefore prevents unwanted surfaces from being formed. In the case of entirely sampled surfaces and a correctly computed crust, two neighboring voxel corners never have different labels. However, if the scene surface is not sampled entirely, such segment borders occur even for correct segmentations (see Figure 4.4d). This forces the surface to pass through the two involved voxel corners which, unlike the rest of the surface reconstruction, does not depend on the confidence values. This fixation does not affect the surface in sampled regions, though. We exploit this constraint on the reconstructed surface in our refinement step where we reconstruct particular areas on higher resolution (see Section 4.7).

4.5 Global Confidence Map

The *global confidence map* (GCM) is a mapping $\Gamma : \mathbb{R}^3 \to \mathbb{R}$ that assigns a confidence value to each point in the volume. Our intuition is that each sample point spreads its confidence over a region in space whose extent depends on the sample footprint. Thus, sample points with a small footprint create a focused spot whereas sample points with a large footprint create a blurry blob (see Figure 4.3b). We model the spatial uncertainty of a sample point as a Gaussian γ_s centered at the sample point's position with standard deviation equal to half the footprint size. If the sample points are associated with confidence values we scale the Gaussian accordingly. The *local confidence map* (LCM) γ_s determines the amount of confidence added by a particular sample point s.

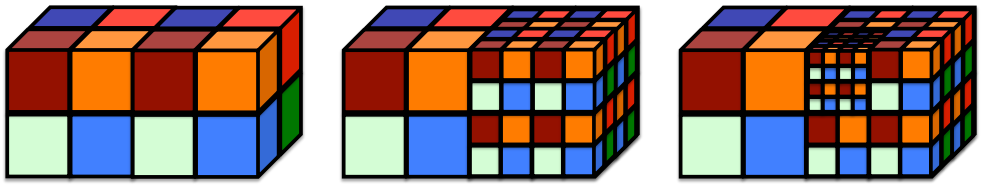


Figure 4.6: Visualization of an intermediate state of the binning approach used for the parallelization of the GCM computation. Starting with two bins (*left*), the right bin is subdivided into eight new bins (*middle*). One of the new bins is subdivided again (*right*) resulting in a total number of 16 bins.

Consequently, the GCM is the sum over all LCMs:

$$\Gamma(x) = \sum_s \gamma_s(x). \tag{4.4}$$

Implementation Let ℓ be the octree level at which we want to compute the graph cut. In all crust voxels $\{x_v\}_{v \in V^\ell_{crust}}$ we evluate the GCM Γ at 27 positions: at the 8 corners of the voxel, at the middle of each face and edge, and at the center of the voxel. When adding up the LCMs of each sample point s we clamp the value of γ_s to zero for points for which the distance to s is larger than three times the footprint size of sample point s. Also, we sample each γ_s only at a fixed number of positions ($\approx 5^3$) within its spatial support and exploit the octree data structure by accumulating each γ_s to nodes at the appropriate octree level depending on the footprint size. After all samples have been processed, the accumulated values in the octree are propagated to the nodes at level ℓ by adding the values at a node to the children's nodes using linear interpolation for in-between positions. The support of LCMs of sample points with small footprints might be too narrow to be adequately sampled on octree level ℓ. For those samples we temporarily increase the footprint for the computation of the LCM γ_s and mark the corresponding voxel for later processing at higher resolution.

4.5.1 Parallelization

In order to speed-up the sample insertion into the octree which is costly since each input point creates ≈ 125 samples, we parallelize the insertion at each octree level $\hat{\ell} \leq \ell$ using a binning approach. In our implementation, bins correspond to voxels. In each bin we sort the samples into eight lists representing the eight child voxels in a predefined order. We process the first list of all bins in parallel, then the second list, and so on. For this purpose samples in list x of two different bins should not interfere with each other, i.e., affect the same nodes in the octree. We start with the bounding cube as root bin

containing all samples to be processed on level $\hat{\ell}$. We subdivide a bin if the following two criteria are satisfied:

1. the bin contains more than n_{max} samples, and
2. subdividing the bin maintains the property that samples out of the same list but different bins do not interfere with each other given their footprint.

When subdividing a bin the lists are effectively turned into bins and the samples are partitioned into eight smaller lists according to the same predefined order as before. The subdivision stops if a maximum number of bins has been reached or no more bins can be subdivided. Figure 4.6 shows the main principle of the subdivision process where the color coded voxels represent the individual lists. Note that two voxels with the same color never touch so that the LCM of samples do not interfere with each other.

4.6 Graph Cut

As done by Hornung and Kobbelt [2006b] we apply a graph cut to find the optimal surface. The layout of the graph cut is however more similar to Boykov and Kolmogorov [2003] since we define a graph node per voxel and edges representing the 26-neighborhood (inside the set of crust voxels V_{crust}). Note that at this stage we compute the graph cut on a certain resolution only and do not extract the surface explicitly. The edge weights w_i in the graph are derived from the GCM values $\Gamma(x_i)$ in the center of the voxel, edge, or face, respectively. Since the optimal surface should maximize the global confidence Γ we want to set small edge weights for regions with high confidence and vice versa. A straightforward way to implement this would be

$$w_i = 1 - \frac{\Gamma(x_i)}{\Gamma_{max}} + a \quad \text{with} \quad \Gamma_{max} = \max_{x \in \mathbb{R}^3} \Gamma(x) \tag{4.5}$$

such that all edge weights lie in $[a, 1+a]$, where a controls the surface tension. Note, that scaling all edge weights with a constant factor does not change the resulting set of cut edges. As the global maximum Γ_{max} can be arbitrarily large, local fluctuation of the GCM might be vanishingly small in relation to Γ_{max} (see Figure 4.7 left). Since the graph cut also minimizes the surface area while maximizing for confidence, the edge weights need to have sufficient local variation to avoid that the graph cut only minimizes the number of cut edges and thus the surface area (*shrinking bias*). In order to cope with that, we apply a technique similar to an adaptive histogram equalization which we call *local GCM balancing*. Instead of using the global maximum in Equation 4.5 we replace

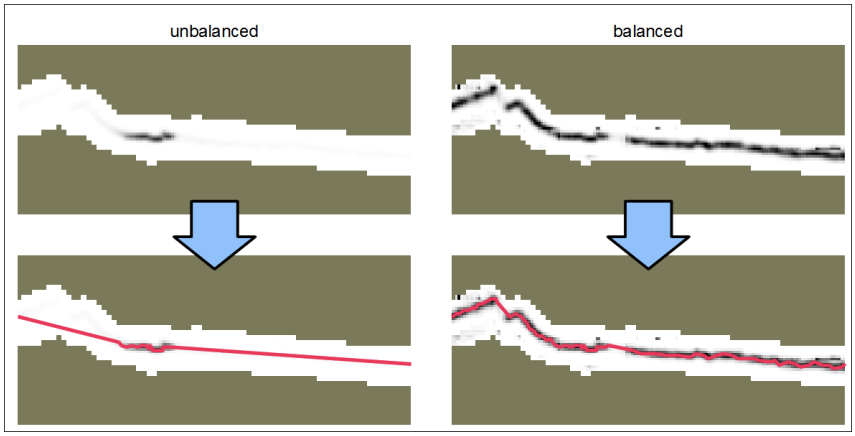


Figure 4.7: The GCM values can be arbitrarily large leading to near-constant edge weights in large regions of the volume (*left*). Our *local GCM balancing* compensates for that allowing the final graph cut to find the correct surface (*right*).

it with the weighted local maximum (LM) of the GCM at point x. We compute $\Gamma_{LM}(x)$ by

$$\Gamma_{LM}(x) = \max_{y \in \mathbb{R}^3} \left[W\left(\frac{\|x-y\|}{2^{-\ell} \cdot \mathscr{B}_{edge}} \right) \cdot \Gamma(y) \right] \tag{4.6}$$

where $\mathscr{B}_{edge}$ is the edge length of the bounding cube. We employ a weighting function W to define the scope in which the maximum is computed. We define W as

$$W(d) = \begin{cases} 1 - \left(\frac{d}{\frac{1}{2}\mathscr{D}} \right)^c & \text{if } d \leq \frac{1}{2}\mathscr{D} \\ 0 & \text{if } d > \frac{1}{2}\mathscr{D} \end{cases} \tag{4.7}$$

where $\mathscr{D}$ is the filter diameter in voxels. We used $\mathscr{D} = 11$ and $c = 4$ in all our experiments. W is continuous in order to ensure continuity of the GCM. See Figure 4.7 (right) to see the effect of local GCM balancing.

After the graph cut, each voxel corner on octree level ℓ is either labeled interior or exterior which we can think of as binary signed distance values. In particular, since the subdivision from level $\ell - 1$ is regular we have labels for all voxel corners, the voxel center, the center of each face and edge. This will be exploited during final surface extraction in the next Section.

4.7 Multi-Resolution Surface Reconstruction

Due to memory limitations, it is often impossible to reconstruct the whole scene on a resolution high enough to capture all sampled details. An adaptive multi-resolution

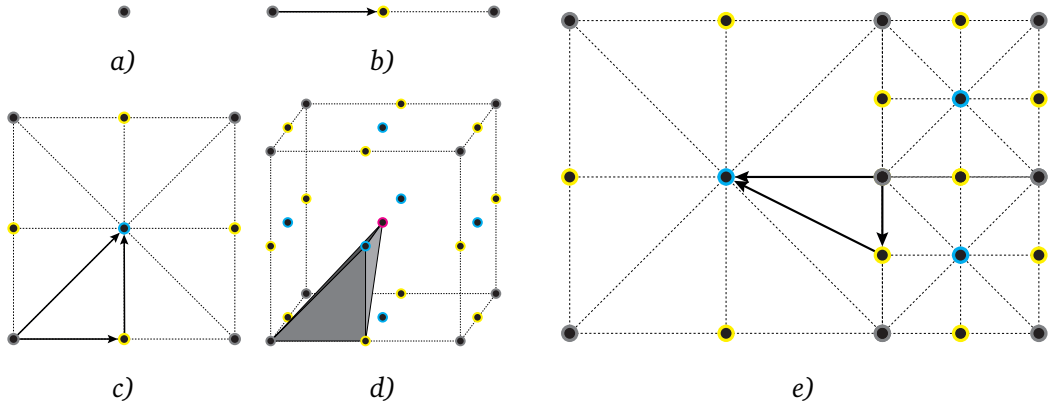


Figure 4.8: Tetrahedralization of the multi-resolution grid. We connect a vertex (a) with the dual vertex of an edge (b), add a face vertex (c), and form a tetrahedron by adding the dual vertex of a cell (d). Adaptive triangulation of the multi-resolution grid (e). Tetrahedralization scheme and figures similar to Manson and Schaefer [2010].

approach which reconstructs different scene regions on adaptive resolutions depending on the sample footprints is therefore desirable. During the GCM sampling on octree level ℓ we marked voxels that need to be processed on higher resolution. After the graph cut we dilate this set of voxels several times and regularly subdivide the resulting voxel set to obtain a new crust $V_{crust}^{\ell+1}$. The crust segmentation can be obtained from the graph cut on level ℓ, as this cut effectively assigns each voxel corner a label *interior* or *exterior*. For boundary voxel corners in $V_{crust}^{\ell+1}$ that coincide with voxel corners on level ℓ we simply transfer the label. This ensures a continuous reconstruction across level boundaries. For voxel corners that lie on a parent voxel edge or face, i.e., between two or four voxel corners on level ℓ, we obtain the conform label of the surrounding voxel corners or we leave it unknown. The new crust $V_{crust}^{\ell+1}$ is now ready for graph cut optimization on level $\ell+1$ (see Figure 4.3d+e). For voxel corners that coincide with voxel corners on the lower resolution the resulting labeling on level $\ell+1$ overwrites the labeling obtained before.

The recursive refinement stops if the maximum level ℓ_{max} is reached or no voxels are marked for further processing. Due to our refinement scheme the last subdivision in the octree is always regular, i.e., all eight octants are present. The graph cuts define the voxel corners of the finest voxels as interior or exterior.

4.7.1 Final Surface Extraction

To extract the final surface we apply a combination of marching cubes and marching tetrahedra. The decision is made voxel-by-voxel one level above the finest level. Note that the last subdivision step is always regular. If the voxel is single-resolution containing 27 labeled voxel corners, we apply classical marching cubes to all eight child voxels. We interpret the voxel corner labels as binary signed distance values. If the voxel is multi-resolution, *i.e.*, there is a change in resolution present affecting at least one of the cube edges or faces, we apply the tetrahedralization scheme by Manson and Schaefer [2010] (see Figure 4.8). We hereby place dual vertices at voxel corners and at the centers of edges, faces, and voxels. These positions coincide with voxel corners of the finest levels providing the binary signed distance values needed for the subsequent marching tetrahedra. Now, we only need to take care of voxel faces where triangles produced by marching cubes and triangles produced by marching tetrahedra meet. It is possible that T-vertices were created here but this can be easily fixed using an edge flip or vertex collapse. The final multi-resolution surface mesh is watertight and has different sized triangles depending on the details present in the corresponding areas.

4.8 Results

We will now present results of our method on different data sets (see Table 4.1). The source code is publicly available on the project page[1]. Our experiments were performed on a 2.7 GHz AMD Opteron with eight quad-core processors and 256GB RAM. All input data was generated from images using a robust structure-from-motion system [Snavely et al. 2008] and an implementation of a recent MVS algorithm [Goesele et al. 2007] applied to down-scaled images. We used all reconstructed points from all depth maps as input samples for our method. The footprint size of a sample is computed as the diameter of a sphere around the sample's 3D position whose projected diameter in the image equals the pixel spacing. For all graph cuts involved we used the publicly available library[2] by Boykov and Kolmogorov [2004].

The Temple is a standard data set provided by the Middlebury Multi-View Stereo Evaluation Project [Seitz et al. 2013] and consists of 312 images showing a temple figurine. This data set can be considered to be single-resolution since all input images have the same resolution and distance to the object, resulting in the complete temple surface to be reconstructed on the same octree level in our algorithm. The reconstruction qual-

[1] http://www.gris.tu-darmstadt.de/projects/multires-surface-recon/
[2] http://vision.csd.uwo.ca/code/

data set	sample points	vertices	octree level	comp. time	rel. variation in footprint
Temple	22 M	0.5 M	9	1 h	1.5
Kopernikus	32 M	3.3 M	10–12	1.5 h	38
Stone	43 M	4.3 M	8–14	4.5 h	75
Citywall	80 M	8.6 M	11–16	6 h	209

Table 4.1: The data sets we used and the number of sample points, the number of vertices in the resulting meshes, octree levels used for surface extraction, computation time and relative variation in footprint size.

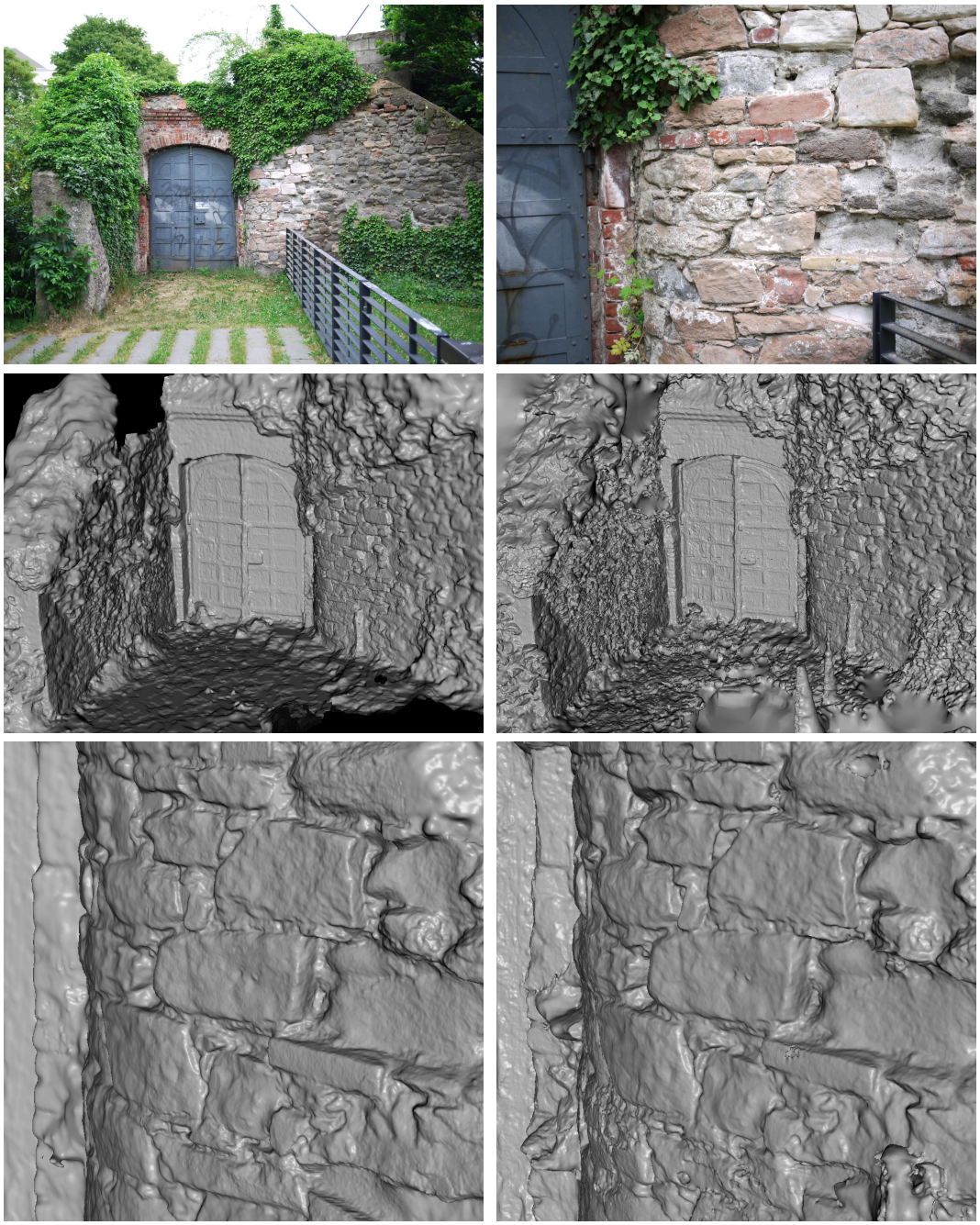
Figure 4.9: An input image of the Temple data set (*left*) and a rendered view of our reconstructed model (*right*).

ity (Figure 4.9) is comparable to other state-of-the-art methods. We submitted reconstructed models created for the VMV paper [Mücke et al. 2011] for the TempleFull and the TempleRing variant (using only a subset of 47 images as input to the pipeline) to the evaluation. For TempleFull we achieved the best accuracy (0.36 mm, 99.7 % completeness), for the TempleRing we achieved 0.46 mm at 99.1 % completeness.

The stone data set consists of 117 views showing a region around a portal where one characteristic stone in the wall is photographed from a close distance leading to high-resolution sample points in this region. Overall we have a factor of 75 of variation in footprint sizes. In Figure 4.10 we compare our reconstruction with Poisson surface reconstruction [Kazhdan et al. 2006]. In the overall view our reconstruction looks significantly better, especially on the ground where our method results in less noise. In the close-up view also Poisson surface reconstruction shows the fine details. Due to the fact that the sampling density is much higher around the particular stone Poisson surface reconstruction used smaller triangles for the reconstruction.

The Citywall data set consists of 487 images showing a large area around a city wall. The wall is sampled with medium resolution, two regions though are sampled with very high resolution: the fountain in the middle and a small sculpture of a city to the left (Figure 4.11 top). Our multi-resolution method is able to reconstruct even fine details in the large scene where sample footprints differ up to a factor of 209. In consequence, the reconstruction spans six octree levels and detailed regions are triangulated about 32 times finer than low-resolution regions. The middle image of Figure 4.11 shows the entire mesh whereas the bottom images show close-ups of the highly detailed surface regions. One can even recognize some windows of the small buildings in the reconstructed geometry.

The Kopernikus data set (Figure 4.12) consists of 334 images showing a statue with a man and a women. The underlying surface geometry is particularly challenging due to its high genus. The data set is also multi-resolution in the sense that we took close-up views of the area around the hands. We compare our reconstruction against VRIP [Curless and Levoy 1996] and the depth map fusion by Fuhrmann and Goesele [2011] (Figure 4.13). It is clearly visible that our model contains significantly less noise and shows no clutter around the real surface. Also, the complex topology of the object is captured very well in comparison to the other methods. However, in regions with low-resolution geometry staircase artifacts are visible due to the surface extraction from a binary signed distance field. This is also visible in the wireframe rendering in Figure 4.12 (bottom right) showing the dense triangulation of the women's face versus the coarse triangulation of the men's upper body.



Our results | Poisson surface reconstruction

Figure 4.10: *Top:* Example input images of the stone data set. *Middle + Bottom:* Comparison of our reconstruction (left) with Poisson surface reconstruction [Kazhdan et al. 2006] (right). Although Poisson surface reconstruction does not take footprints into account the reconstruction shows fine details due to the higher sampling density. However, our surface shows significantly less noise and clutter.

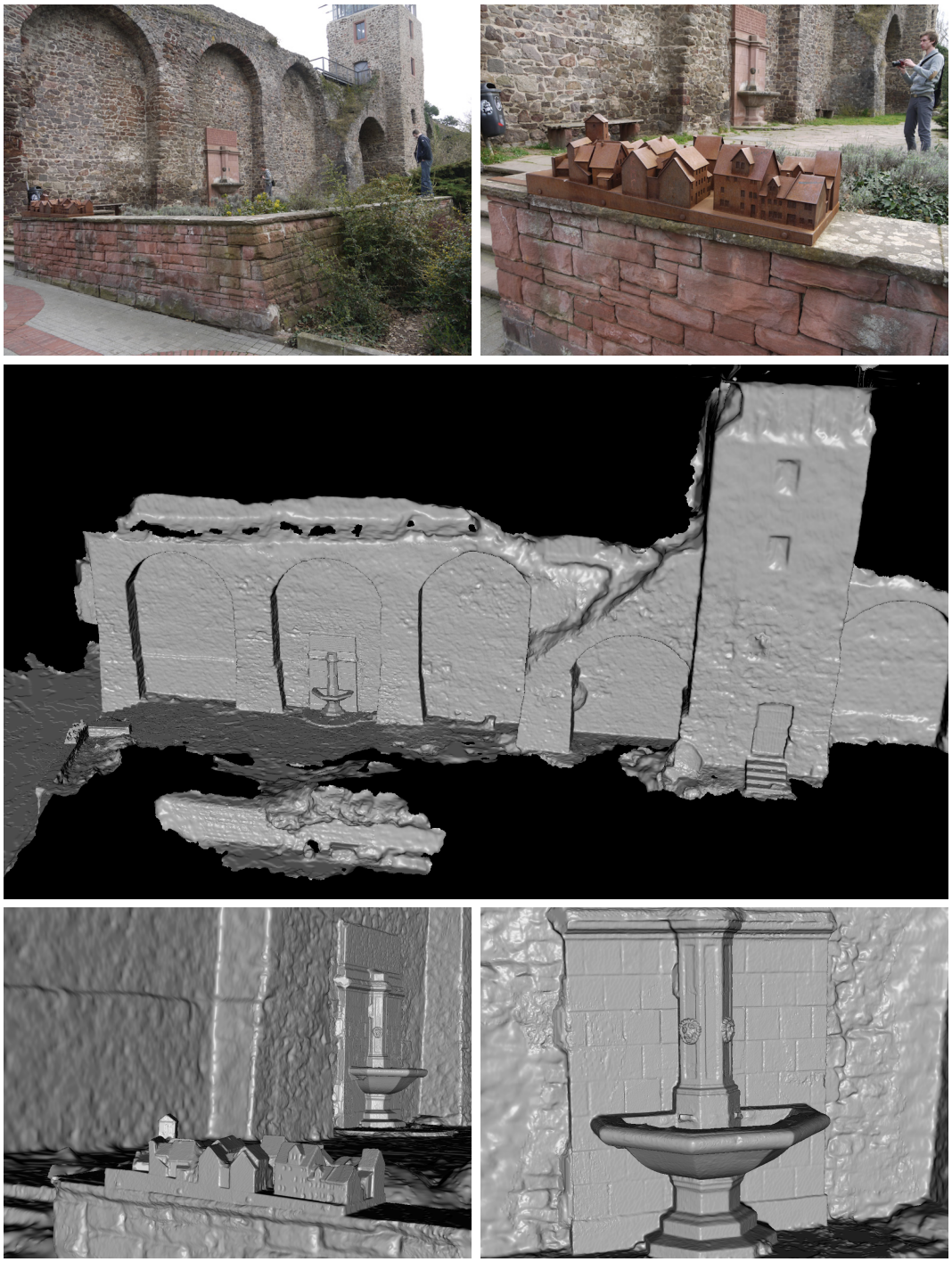
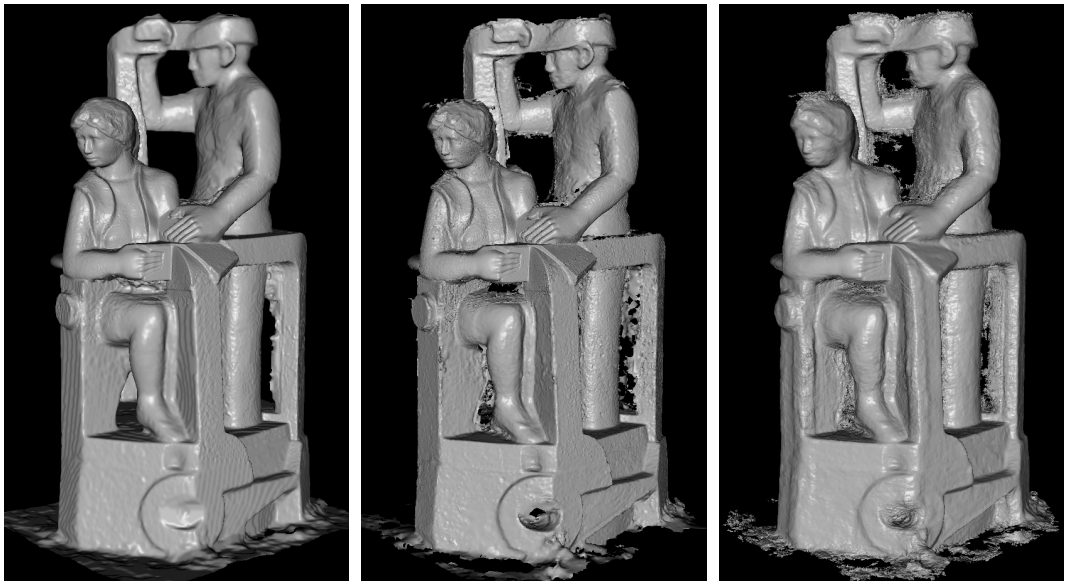


Figure 4.11: *Top:* Two input images of the Citywall data set. Middle: Entire model (color indicates the octree level, red is highest). *Bottom:* Close-ups of the two detailed regions.

Figure 4.12: Two input images of the Kopernikus data set, the complete reconstructed model from two perspectives and a close-up of the wireframe showing the adaptively triangulated mesh.



Our results | Depth map fusion | VRIP

Figure 4.13: Comparison of our reconstruction (*left*) with depth map fusion [Fuhrmann and Goesele 2011] (*middle*) and VRIP [Curless and Levoy 1996] (*right*).

4.9 Discussion

We presented a robust surface reconstruction algorithm that works on general input data. To our knowledge, except for the concurrent work of Fuhrmann and Goesele [2011], we are the first to take the footprint of a sample point into account during reconstruction. Together with a robust crust computation and an adaptive multi-resolution reconstruction approach we are able to reconstruct fine detail in large-scale scenes. We presented results comparable to state-of-the-art techniques on a benchmark data set and proved our superiority on challenging large-scale outdoor data sets and objects with complex topology. The triangle meshes are manifold and watertight and show an adaptive triangulation with smaller triangles in regions with higher details.

Future work includes to explore other ways to distribute a sample point's confidence over the volume, e.g., taking the direction to the sensor into account. This allows for better modeling the generally anisotropic error present in reconstructed depth maps.

5 Modulation Transfer Function of Patch-based Stereo Systems

Contents

A widely used technique to recover a 3D surface from photographs is patch-based (multi-view) stereo reconstruction. Current methods are able to reproduce fine surface details. They are however limited by the sampling density and the patch size used for reconstruction. We show that there is a systematic error in the reconstruction depending on the details in the unknown surface (frequencies) and the reconstruction resolution. For this purpose we present a theoretical analysis of patch-based depth reconstruction. We prove that our model of the reconstruction process yields a linear system, allowing us to apply the transfer (or system) function concept. We derive the modulation transfer function theoretically and validate it experimentally on synthetic examples using rendered images as well as on photographs of a 3D test target. Our

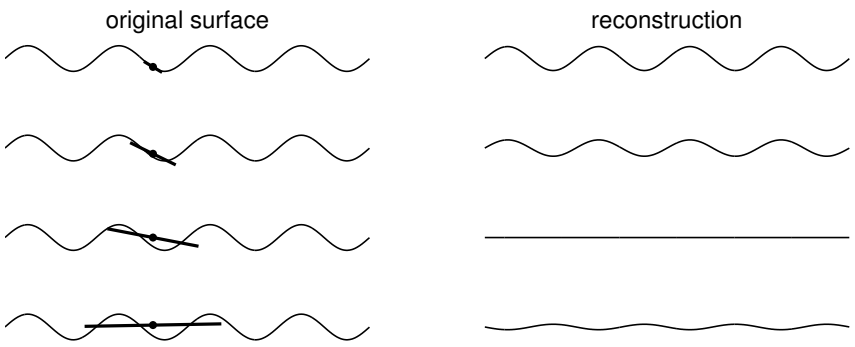
analysis proves that there is a significant but predictable amplitude loss in reconstructions of fine scale details. In a first experiment on real-world data we show how this can be compensated for within the limits of noise and reconstruction accuracy by an inverse transfer function in frequency space.

5.1 Introduction

Patch-based (multi-view) stereo reconstruction [Bradley et al. 2008, Furukawa and Ponce 2010, Goesele et al. 2007, Habbecke and Kobbelt 2007, Jancosek et al. 2009] is a widely used technique to recover a 3D surface from photographs. Current methods achieve remarkable accuracy and are able to capture even fine geometric details [Seitz et al. 2006]. Their ability to faithfully reconstruct details is obviously limited by two facts: the sampling density of the algorithm and the size of the patch used for reconstruction (both of these are typically coupled to the resolution of the input images). To give a concrete example: a planar surface modulated with fine scale detail will eventually be reconstructed as a plane as image resolution decreases and patch size increases. This is illustrated in Figure 5.1 for a 1D signal.

We are interested in the geometry reconstructed by a patch-based algorithm for details that are roughly at the scale of the patch size. As also illustrated in Figure 5.1, such details are *reconstructed with much lower amplitude* and can even be inverted, so that valleys are reconstructed as peaks and vice versa. This behavior is not only contradicting our standard (or naïve) intuition about the properties of patch-based reconstruction, it is also in stark contrast to the assumptions made by most fusion techniques used to reconstruct a single surface from a set of reconstructed points or depth maps. These algorithms typically assume that the reconstructed points are samples of the true surface disturbed by zero-mean Gaussian noise [Curless and Levoy 1996, Kazhdan et al. 2006, Zach et al. 2007]. Different scales or sampling densities are sometimes represented by lower confidences (or large variances in the noise model) and often enough just ignored. This implies that a reliable measurement of the true surface can be obtained by just averaging enough surface samples as this will cancel out noise.

In this chapter, we show that there is a *systematic error* in the reconstruction depending on the details in the unknown surface (frequencies) and the reconstruction resolution. We show that even a "perfect" patch-based reconstruction algorithm will result in different reconstructed geometry of the same scene if used at different scales (e.g., varying resolution of input images or changing patch size). To our knowledge this fact is not modeled in any existing patch-based reconstruction algorithm. We provide



wavelength unit: patch width	**5**	**2**	$\frac{3}{2}$	**1**	$\frac{2}{3}$	$\frac{1}{2}$
relative amplitude of reconstruction	**0.94**	**0.64**	0.41	**0**	**-0.21**	0

Figure 5.1: Predicted reconstruction of a sinusoidal surface with different patch widths. *Top:* The amplitude of the reconstruction varies drastically with the width of the patch used for reconstruction. In some cases, the signal is even inverted. The bold line marks the optimal patch position and orientation. *Bottom:* Table with predicted amplitude loss depending on patch width relative to signal wave length. Bold columns mark the cases drawn above.

a model that predicts how amplitudes of different frequencies in the incoming signal are reproduced. The model is motivated by the concept of optical transfer functions (OTF) [Szeliski 2010, Williams 1999] typically applied in the context of 2D image processing. It allows us theoretically to *invert this process*, in practice however only within the limits of noise and reconstruction accuracy.

The remainder of this chapter is organized as follows: We first review related work (Section 5.2) before we derive and validate our model in 2D using synthetic examples and a real-world test target (Section 5.3). We then extend our theory to 3D (Section 5.4) and show its relevance on a real life application. Finally, we discuss our results (Section 5.5).

5.2 Related Work

The analysis of different scale geometry reconstruction using patch-based stereo techniques has been neglected so far. For an overview and classification of multi-view stereo we refer to the recent survey [Seitz et al. 2006] and constantly updated benchmark [Seitz et al. 2013]. Key elements in our work build upon signal processing, optical transfer functions, and multi-scale surface representation. Existing work of the latter

two areas will be discussed in the following.

The *optical transfer function* (OTF) is a well known concept to describe how details are reproduced by an imaging system [Williams 1999]. It relies on the assumption of a linear system and describes how amplitude and phase change for different frequencies in the image using modulation and phase transfer functions, respectively. In our work, we validate that the linearity assumption holds and estimate the modulation transfer function of a patch-based stereo system. The OTF can be estimated in various ways [Williams 1999]. For sampled imaging systems, Reichenbach *et al*. [1991] introduced the knife-edge technique. Multiple scan lines are first registered to create a super-resolution edge profile and to suppress noise before the frequency space behavior is analyzed. Goesele *et al*. [2003] applied this technique to estimate the modulation transfer function of a 3D range scanner. They capture a slanted edge and fit two planes to the measurements to create a super-resolution edge profile. The Fourier transform of the profile is then compared to that of an ideal edge.

Kobbelt *et al*. [1998] define *multi-scale surface representations* and encode changes between levels using normal displacements. They use fairing operators to iteratively smooth a mesh and apply the results in the context of multi-scale surface editing. Inspired by Lindeberg's scale-space theory [Lindeberg 1994], Pauly *et al*. [2006] present a point-based multi-scale representation scheme using approximate geometric low-pass filtering and a projection operator to encode the different levels of detail. They discuss two approximate low-pass filters based on diffusion and least squares filtering, respectively. Both can lead to deformations such as surface shrinkage. They identify the problem that no global, distortion-free parameterization exists for manifolds in general.

In this chapter, we draw the *connection* between multi-scale surface representations and patch-based stereo reconstruction. We rely on the transfer function concept and the analysis techniques presented above, allowing us to demonstrate the effects in theory and practice. Using the simplifying assumption that the geometry can be represented as a height field, we are able to apply Fourier analysis to the reconstructed geometry.

5.3 Modeling the Reconstruction Process

The common strategy in patch-based stereo methods is to locally fit a planar patch to the unknown geometry that is photo-consistent with one or more other views. A typical example for measuring photo-consistency is the normalized cross-correlation (NCC) of points on the patch projected in other views. The final surface is represented by the (triangulated) central patch points [Bradley et al. 2008, Furukawa and Ponce

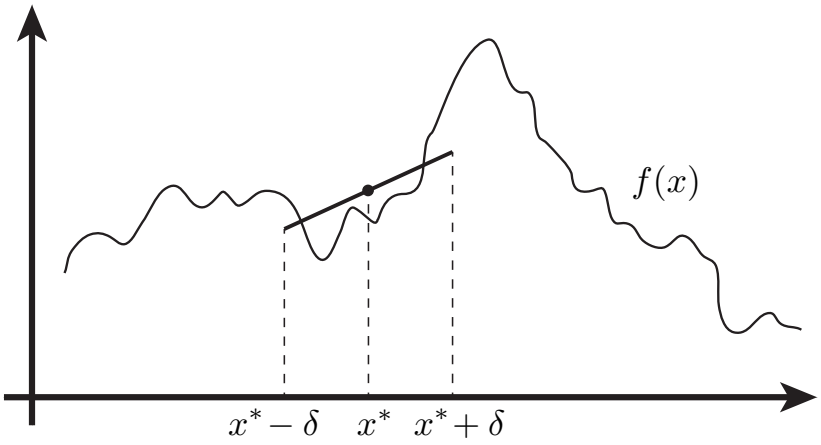


Figure 5.2: Fitting a planar patch (line segment) to the geometry for each point x^*.

2010, Habbecke and Kobbelt 2007] or the points are merged into a distance field [Curless and Levoy 1996, Fuhrmann and Goesele 2011, Zach et al. 2007]. In the following, we will develop a theoretical model for fitting a planar patch to the geometry, first in 2D and later in 3D (Section 5.4).

We assume that the geometry can be described as a height field $z = f(x)$ (i.e., the whole surface is visible from an orthographic camera aimed perpendicular to the height field plane). In order to obtain the reconstruction $\hat{z} = \hat{f}(x)$ at position x^* we fit a patch (line segment) with an extent of 2δ centered around x^* to the geometry. Figure 5.2 visualizes the idea for a 2D geometry. We represent the line segment by two parameters m, n and model the fitting process as optimizing for least squares distance to the true geometry by minimizing the following energy

$$E(m, n, x^*) = \int_{x^*-\delta}^{x^*+\delta} (mx + n - f(x))^2 dx. \tag{5.1}$$

The reconstructed surface height at x^* is then given through the optimal parameters m, n by $\hat{z} = mx^* + n$. Note that we measure the patch extent along the x-axis in world coordinates and not in pixels as typically done in stereo. In the remainder of the chapter we will use the term *patch width* for describing a patch of extent 2δ. The parameter δ also depends on image resolution, surface distance to the camera, and the camera's focal length. The actual patch size depends however on the slope (or orientation) of the patch. Intuitively, a smaller δ allows to capture fine details whereas a larger δ yields a smoothed surface. Image resolution often defines the sampling frequency equal to the distance between two consecutive points x_1^* and x_2^* where we fit a patch. In the

following, we will deliberately disregard image resolution and think of reconstructing the geometry as fitting a patch continuously at every point x^*.

5.3.1 Theoretical Results for a Sine Wave

We start by analyzing the simplest geometry in the sense of frequency behavior, a sine wave $f(x) = a\sin(\omega x)$ with amplitude a and frequency ω. To determine the reconstructed signal according to our model, we need to minimize E by finding the roots of the partial derivatives

$$\partial_m E = 2\int_{x^*-\delta}^{x^*+\delta} x(mx+n-a\sin(\omega x))dx \stackrel{!}{=} 0 \tag{5.2}$$

$$\partial_n E = 2\int_{x^*-\delta}^{x^*+\delta} (mx+n-a\sin(\omega x))dx \stackrel{!}{=} 0. \tag{5.3}$$

Solving the equations for m and n results in

$$m = \frac{3a\cos(\omega x^*)(\sin(\omega\delta)-\omega\delta\cos(\omega\delta))}{\omega^2\delta^3} \tag{5.4}$$

$$n = \frac{a\delta^2\omega\sin(\omega x^*)\sin(\omega\delta)}{\omega^2\delta^3} + \frac{3ax^*\cos(\omega x^*)(\omega\delta\cos(\omega\delta)-\sin(\omega\delta))}{\omega^2\delta^3} \tag{5.5}$$

Inserting this in $\hat{z} = mx^* + n$, the reconstruction is

$$\hat{f}(x^*) = \frac{a\sin(\omega\delta)\sin(\omega x^*)}{\omega\delta} = a\,\mathrm{sinc}(\omega\delta)\sin(\omega x^*). \tag{5.6}$$

This is an interesting result because frequency and phase of the sine are preserved for arbitrary patch width and frequency; only the amplitude is scaled by $\mathrm{sinc}(\omega\delta)$ confirming one part of our linear system assumption. Note that for certain combinations $\omega\delta$ the signal can even be inverted so that valleys become peaks and vice versa. In the following we will corroborate this result experimentally.

5.3.2 Experimental Results for a Sine Wave

We first validate our results on synthetic data sets, rendered using the PBRT system[1]. This has the advantage that registration is perfect and all observed effects are due to photo-consistency optimization alone. As test target, we create a mesh representing a sine wave in the x,y—plane with $z(x,y) = a\sin(\omega x)$. The mesh is observed by five

[1] http://www.pbrt.org

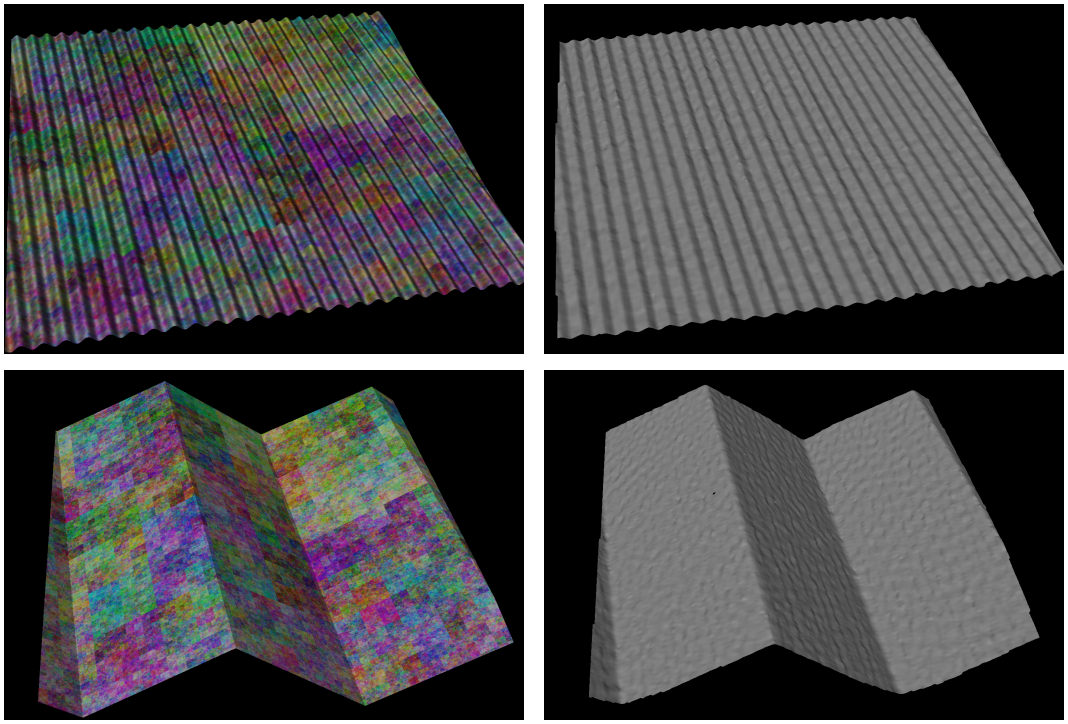


Figure 5.3: *Left:* Screenshot of the textured meshes used for our synthetic experiments. *Right:* Sample multi-view stereo reconstructions.

perspective cameras: One central camera points orthogonal to the x, y—plane and the other cameras are equally distributed around it with 15° parallax. A random texture with structure on all scales is mapped onto the geometry (see Figure 5.3(*left*)). We render views of the geometry using a variety of image resolutions. For the highest resolution we also create a ground truth depth map. For reconstruction, we run a patch optimization taken from an existing multi-view stereo system [Goesele et al. 2007, Sect. 6.2] using the central camera as reference view and the surrounding cameras as neighbor views. For each pixel in the central camera the optimization is initialized with a fronto-parallel patch at depth values associated with that pixel in the highest-resolution ground truth depth map. The optimized patch with highest confidence (based on NCC) determines the depth at the current pixel. See Figure 5.3(*right*) for example reconstructions for images of resolution 256×256 with image patch size 5×5 pixel. Note the regular structure introduced by the strong texture gradients most notably in the zigzag shape.

For data analysis, we fit the parameters amplitude $\hat{a}$, frequency $\hat{\omega}$, phase $\hat{p}$ and offset $\hat{o}$ of the sine function $z = \hat{a}\sin(\hat{\omega}x + \hat{p}) + \hat{o}$ to all reconstructed 3D points using

Levenberg-Marquardt optimization[2]. To obtain a super-resolution sampling of the sine wave along the x-axis the camera's up-vector is slightly tilted against the y-axis (about 5°) similar to the knife edge technique [Reichenbach et al. 1991]. In our experiments we use two sine waves of different frequency ($\omega = 32$ and $\omega = 64$). We vary the patch width parameter δ by using various image resolution as well as image patch sizes of 5×5 and 7×7 pixels. Figure 5.4 shows that the reconstructed relative amplitudes, relative frequencies, phases, and offsets match very well with the predicted values. The observed differences are primarily caused by imperfections in the reconstruction process, in particular the interaction between the model texture and the photo-consistency of the patch.

5.3.3 Stereo Transfer Function

Ideally, we can express the reconstruction process using a transfer (or system) function representing the relation between input and output in terms of spatial frequencies. This concept is common in the imaging domain (optical transfer function) [Szeliski 2010, Williams 1999] for describing the capability of showing fine details and the trade-off between blurred structure and aliasing. The optical transfer function is actually the Fourier transform of the point spread function. However, the transfer function concept is only applicable to linear systems featuring the principle of superposition and stationarity. The latter is given for our model since the reconstruction is lateral shift invariant. What remains to check is the principle of superposition or additivity. We show that if the geometry is the sum of different frequency components the reconstruction is the sum of its separate contributions. For this purpose we represent f by a complete Fourier series

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos(kx) + b_k \sin(kx)). \tag{5.7}$$

Again, we need to find m and n, so that on the interval $I = [x^* - \delta, x^* + \delta]$ the energy $E(m, n, x^*)$ is minimized:

$$E(m, n, x^*) = \int_I (mx + n - f(x))^2 \, dx. \tag{5.8}$$

[2] http://www.ics.forth.gr/~lourakis/levmar/

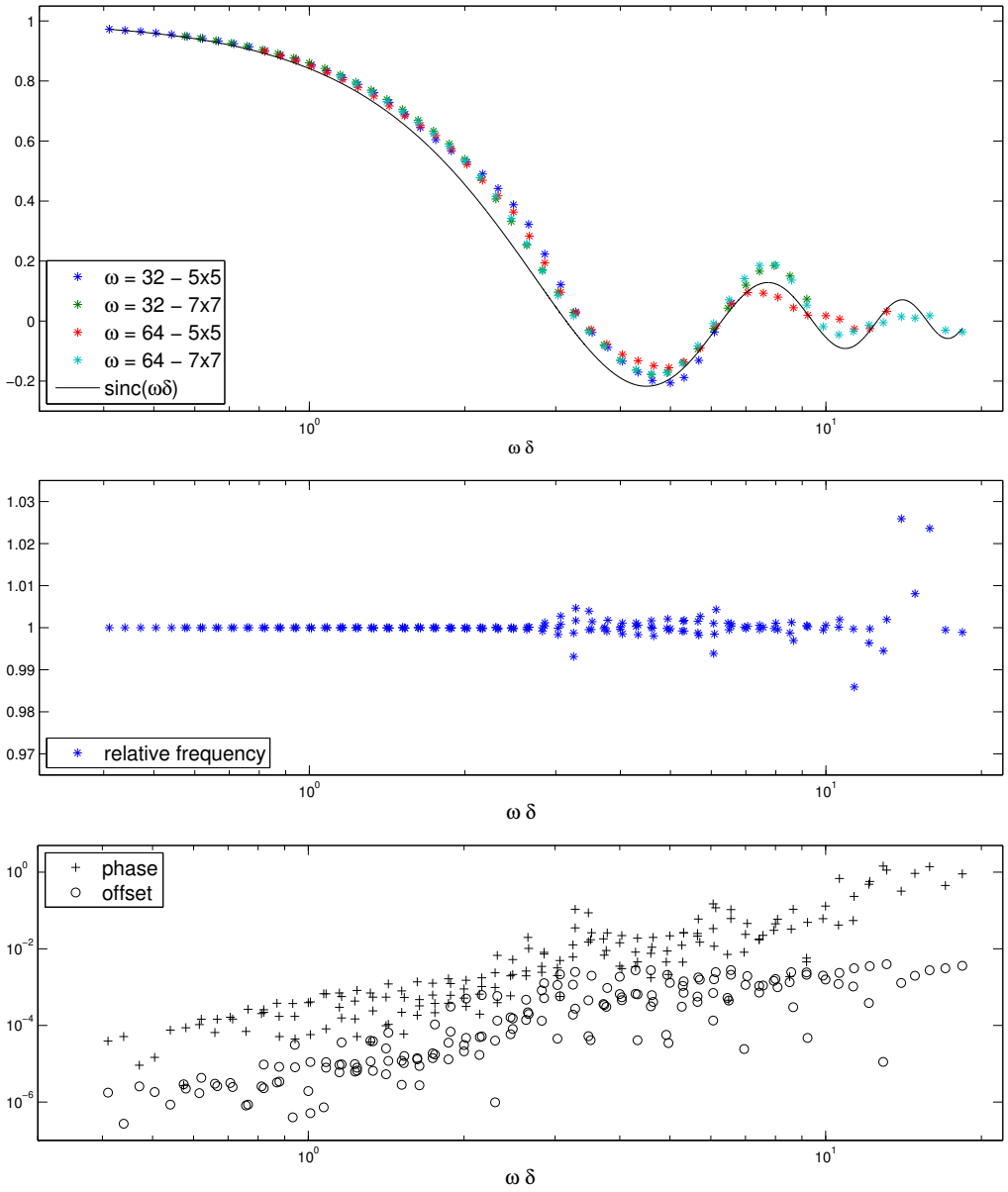


Figure 5.4: Resulting relative amplitude, relative frequency, phase, and offset of the reconstructed sine wave for different wavelengths and patch widths.

This implies taking partial derivatives with respect to m and n and finding the roots of these equations:

$$\begin{aligned}\partial_m E(m,n,x^*) &= \int_I 2x\,(mx+n-f(x))\,dx \stackrel{!}{=} 0\\ \partial_n E(m,n,x^*) &= \int_I 2\,(mx+n-f(x))\,dx \stackrel{!}{=} 0.\end{aligned} \tag{5.9}$$

This yields the following solution for Equation 5.9:

$$\begin{aligned}E_m &= nx^2+\frac{2}{3}mx^3-\frac{1}{2}a_0x^2+\sum_{k=1}^{\infty}\frac{2}{k^2}\,(-a_k\cos(kx)\\ &\quad -b_k\sin(kx)-kya_k\sin(kx)+kyb_k\cos(kx))\\ E_n &= 2nx+mx^2-xa_0\\ &\quad -\sum_{k=1}^{\infty}\frac{2}{k}\,(a_k\sin(kx)-b_k\cos(kx))\,.\end{aligned} \tag{5.10}$$

Inserting the boundaries of the interval I (ignoring the superscript * for typographic reasons) in Equation 5.10 yields

$$\begin{aligned}0 &= 4nx\delta+4mx^2\delta+\frac{4}{3}m\delta^3-2x\delta a_0+\sum_{k=1}^{\infty}\frac{4}{k^2}\,(\\ &\quad -xka_k\cos(kx)\sin(k\delta)-\delta ka_k\sin(kx)\cos(k\delta)\\ &\quad +a_k\sin(kx)\sin(k\delta)-xkb_k\sin(kx)\sin(k\delta)\\ &\quad +\delta kb_k\cos(kx)\cos(k\delta)-b_k\cos(kx)\sin(k\delta))\\ 0 &= \Big(-4\delta mx-4\delta n+2\delta a_0+\\ &\quad \sum_{k=1}^{\infty}\frac{4}{k}\sin(k\delta)\,(a_k\cos(kx)+b_k\sin(kx))\Big)\end{aligned} \tag{5.11}$$

These two equations are linear in m and n and can be easily solved. Moreover, from Equation 5.11 one obtains the expression for the solution $mx+n$ (the reconstructed geometry) directly as

$$\hat{f}(x)=\frac{a_0}{2}+\sum_{k=1}^{\infty}\operatorname{sinc}(k\delta)\,(a_k\cos(kx)+b_k\sin(kx))\,. \tag{5.12}$$

Thus, the principle of superposition is fulfilled and our model of patch-based stereo reconstruction is a linear system. This allows us to formulate the relationship between reconstructed and real geometry as

$$\hat{F}_\delta(\omega)=\mathrm{MTF}_\delta(\omega)\cdot F(\omega)=\operatorname{sinc}(\omega\delta)\cdot F(\omega) \tag{5.13}$$

where $\hat{F}_\delta$ and F are the Fourier transforms of the reconstructed (using patch width 2δ) and real geometry. $\mathrm{MTF}_\delta(\omega)$ is the modulation transfer function. Note that there is a difference to the traditional OTF. In our case the MTF can also be negative, modeling an inversion of amplitudes and the geometry, respectively. This allows us to completely remove the phase transfer function. In the next section, we will validate this result experimentally.

5.3.4 Experiments on a Slanted Edge

To experimentally validate Equation 5.13 we reconstruct a zigzag shape whose Fourier transform contains frequencies on all scales due to its sharp edges. Apart from the underlying geometry which is a zigzag shape (constant along y-axis) with edges of about 126° we use the same setup as in Section 5.3.2. Again, we just look at the (x, z) pairs of all reconstructed points. The slanted edge (implemented by the slightly tilted up-vector) gives us a fine sampling of the edge along the x-axis. We chose an interval $[x_{min}, x_{max}]$ such that it captures exactly one period of the zigzag shape and sample all points therein into 2^n bins so that the Discrete Fourier Transform (DFT) can be applied. In the Fourier transform of the ground truth profile every second coefficient is zero so we only use every second coefficient to compute the MTF, where the different resolutions lead to various patch widths 2δ (see Figure 5.5 top, middle). We can also measure how the amplitude is altered according to the product of frequency ω and δ (see Figure 5.5 bottom). Up to $\omega\delta \approx 1.5$ the measured data matches very well with the theoretically predicted result. Beyond that point, the MTF still follows the theoretical prediction $\mathrm{sinc}(\omega\delta)$ but is masked by noise introduced by the reconstruction process.

5.3.5 Results on Real-World Data

Our goal is to analyze an object of simple and known 1D geometry to validate our theory with real world data. We therefore created a test target using 3D printing technology (see Figure 5.6). It consists of two periods of a sine wave with wavelength 62.8 mm and amplitude 10.0 mm and an edge with an angle of about 126°. Both are spread over 188.5 mm in width. To provide structure, we mapped the same texture as used in our synthetic experiments on the entire surface. This model was printed using a ZPrinter® 650 which has a printing accuracy of about 0.1 mm according to manufacturer specifications. For our experiments, we took photos with a digital SLR (one central photo looking orthogonal onto the object and several surrounding photos) with three different average camera distances to the object (near: 95 cm, middle: 145 cm, far: 280 cm).

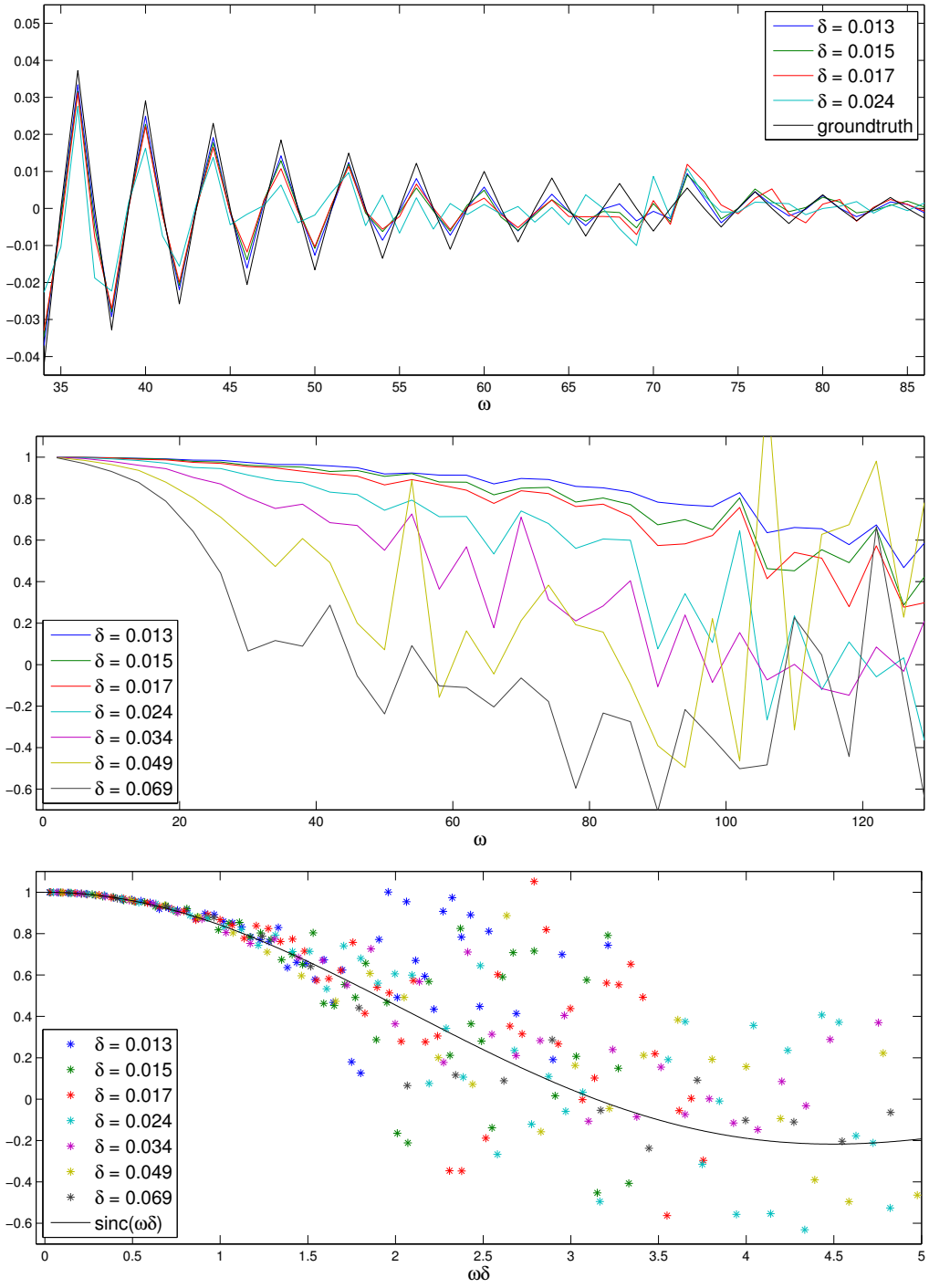


Figure 5.5: *Top:* Imaginary part of DFT coefficients for the zigzag profile. *Middle:* MTF samples for different patch widths 2δ as a function of ω. *Bottom:* MTF as a function of the product $\omega\delta$.

For each set of photos we perform a calibration using structure-from-motion [Snavely et al. 2008]. We then apply a multi-view stereo algorithm [Goesele et al. 2007] with patch-based optimization to compute a depth value for each pixel in the central views. Hereby, we repetitively rescale the images in order to get depth maps of different resolutions and additionally run the reconstruction algorithm with two different image patch sizes (5×5 and 7×7 pixel).

To analyze the amplitude loss on the sine wave, we first determine an optimal transform aligning the reconstruction with the x, y-plane. This optimal transformation is applied to all the different resolution depth maps to which we then fit in a second step a sine with amplitude, frequency, phase, and offset as in our synthetic experiments. Figure 5.7 (top) shows the amplitude loss with growing $\omega\delta$. The results closely match the theoretical prediction. In the second experiment, we analyze the reconstructed edge of the test target using an approach very similar to Goesele *et al.* [Goesele et al. 2003]. We first fit two least squares planes to the (highest resolution) reconstructed points on both sides of the edge and rotate the scan such that the intersection line coincides with the y-axis and the edge profile is symmetric to the y, z—plane. We then bin the reconstructed points ((x, z)-pairs) into 257 bins along the x-axis, move the ends to $z = 0$ and multiply with a Blackman window. Then each profile is rotated around one end point to continue it periodically, dropping the first and last bin and thus resulting in 512 bins. We apply the Fourier transform to each profile and compare it to the Fourier transform of a perfect edge profile. Figure 5.7 (bottom) shows the sampled MTF values for different δ. The result shows significantly more noise and outliers than on the synthetic data reflecting errors in the registration, wrongly matched patches due to far-off start points and summed up errors during region growing.

5.4 Moving from 1D to 2D Functions

So far, we described the theory for one-dimensional functions and validated it using geometry that is constant in one dimension. Naturally, real-world geometry rarely conforms to such a constrained model. We therefore show how our theory extends to height fields parameterized over a 2D plane, i.e., surfaces that can be described by $z = f(x, y)$.

5.4.1 Theory for a Height Field over a 2D Plane

Clearly, the same procedure can be applied in 2D. Let

$$P = m_x x + m_y y + n \tag{5.14}$$

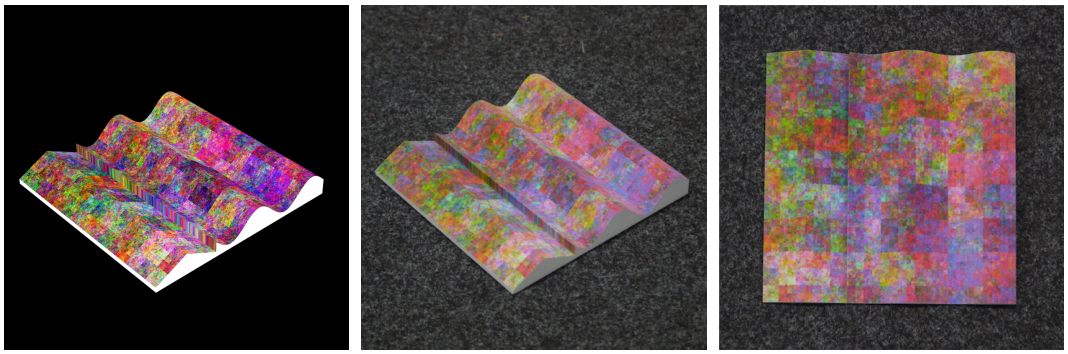


Figure 5.6: *Left:* Rendering of the test target. *Middle/Right:* Side and top view of the manufactured test target.

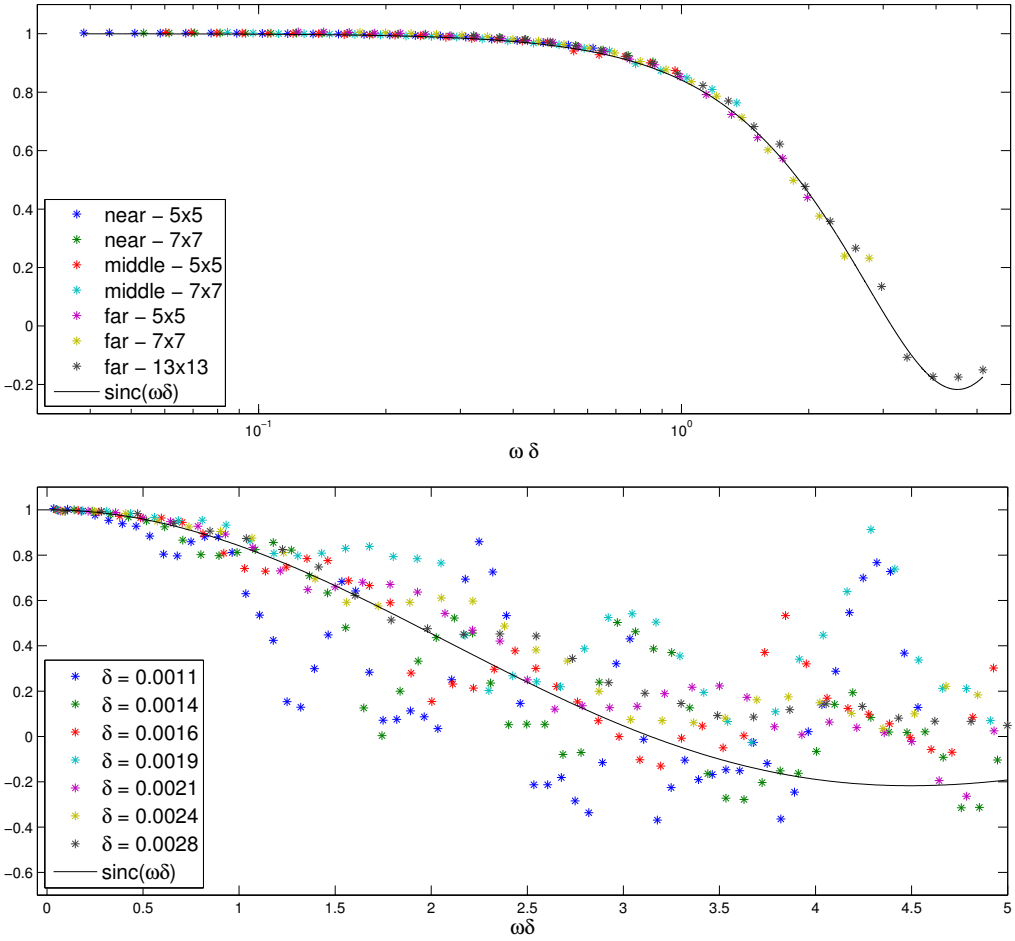


Figure 5.7: Results using the manufactured test target. *Top:* Amplitude loss on the sine wave. *Bottom:* MTF samples using the edge.

be the solution to the patch that we want to compute around a point (x, y) spanned by $I = [x-\delta, x+\delta] \times [y-\epsilon, y+\epsilon]$. Note that this covers the general case of a rectangular patch instead of the usual square patch. The signal $f(x, y)$ can be expressed in terms of a sine and cosine series or, alternatively, using complex numbers by

$$f(x,y) = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \alpha_{j,k} e^{i(jx+ky)} \tag{5.15}$$

Again we want to find the minimum of

$$E = \int_y \int_x (P-f)^2 \, dx \, dy \tag{5.16}$$

for the parameters m_x, m_y, and n. Taking derivatives with respect to these parameters and solving yields

$$E_{m_x} = nx^2y + \frac{2}{3}m_x x^3 y + \frac{1}{2}m_y x^2 y^2 + \sum_{j,k} \alpha_{j,k} e^{i(jx+ky)} \left(\frac{2i}{j^2 k} + \frac{2x}{jk} \right) \tag{5.17}$$

$$E_{m_y} = nxy^2 + \frac{1}{2}m_x x^2 y^2 + \frac{2}{3}m_y x y^3 + \sum_{j,k} \alpha_{j,k} e^{i(jx+ky)} \left(\frac{2i}{jk^2} + \frac{2y}{jk} \right) \tag{5.18}$$

$$E_n = 2nxy + m_x x^2 y + m_y x y^2 + \sum_{j,k} \alpha_{j,k} e^{i(jx+ky)} \frac{2}{jk}. \tag{5.19}$$

On the given patch I we get

$$\begin{aligned} E_{m_x} &= nx + m_x x^2 + m_y xy + \frac{1}{3}m_x \delta^2 \\ &\quad + \sum_{j,k} \alpha_{j,k} e^{i(jx+ky)} \sin(k\epsilon) \cdot \left(\frac{i}{jk\epsilon} \cos(j\delta) - \sin(j\delta)\left(\frac{i}{j^2 k \delta \epsilon} + \frac{x}{jk\delta\epsilon}\right) \right) \end{aligned} \tag{5.20}$$

$$\begin{aligned} E_{m_y} &= ny + m_x xy + m_y y^2 + \frac{1}{3}m_y \epsilon^2 \\ &\quad + \sum_{j,k} \alpha_{j,k} e^{i(jx+ky)} \sin(j\delta) \cdot \left(\frac{i}{jk\delta} \cos(k\epsilon) - \sin(k\epsilon)\left(\frac{i}{jk^2 \delta \epsilon} + \frac{y}{jk\delta\epsilon}\right) \right) \end{aligned} \tag{5.21}$$

$$\begin{aligned} E_n &= n + m_x x + m_y y \\ &\quad - \sum_{j,k} \alpha_{j,k} e^{i(jx+ky)} \frac{1}{jk\delta\epsilon} \sin(j\delta)\sin(k\epsilon) \end{aligned} \tag{5.22}$$

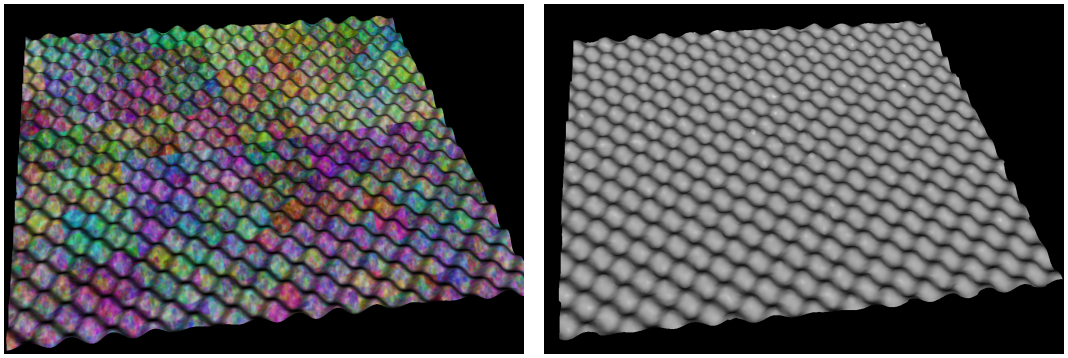


Figure 5.8: *Left:* Screenshot of the textured mesh showing the 2D sine. *Right:* Example reconstruction.

We can solve these linear equations in m_x, m_y, and n. From $E_n = 0$ one can directly derive the solution for our patch:

$$\begin{aligned} P &= m_x x + m_y y + n \\ &= \sum_{j,k} \alpha_{j,k} e^{i(jx+ky)} \frac{1}{jk\delta\epsilon} \sin(j\delta)\sin(k\epsilon) \\ &= \sum_{j,k} \alpha_{j,k} e^{i(jx+ky)} \operatorname{sinc}(j\delta)\operatorname{sinc}(k\epsilon). \end{aligned} \tag{5.23}$$

We see that the amplitude loss is a product of two sinc functions which is the Fourier transform of a box filter.

5.4.2 Results on Synthetic 2D Sine

We will substantiate the theoretical result on geometry containing only one frequency along each dimension and construct a height field with $z = \frac{1}{\omega}\sin(\omega x)\sin(\omega y)$. Figure 5.8 shows a rendering of the textured mesh (*left*) as well as an example multi-view stereo reconstruction for an image of resolution 256×256 pixel with image patch size 5×5 pixel.

Apart from this geometry, the setup is equivalent to that in Sections 5.3.2 and 5.3.4. We optimize for the six parameters amplitude $\hat{a}$, frequencies $\hat{\omega}_x, \hat{\omega}_y$, phases $\hat{p}_x, \hat{p}_y$, and offset $\hat{o}$ such that $z = \hat{a}\sin(\hat{\omega}_x x + \hat{p}_x)\sin(\hat{\omega}_y y + \hat{p}_y) + \hat{o}$ holds for the reconstructed 3D points. According to the theoretical result from Equation **??**, the reconstructed amplitude should be scaled by $\operatorname{sinc}^2(\omega\delta)$ compared to the original amplitude. Figure 5.9 shows that the experimentally obtained scaling factors match the expected values very well. The estimated frequencies, phase shifts, and offsets are comparable to the 1D experiments (similar to Figure 5.4).

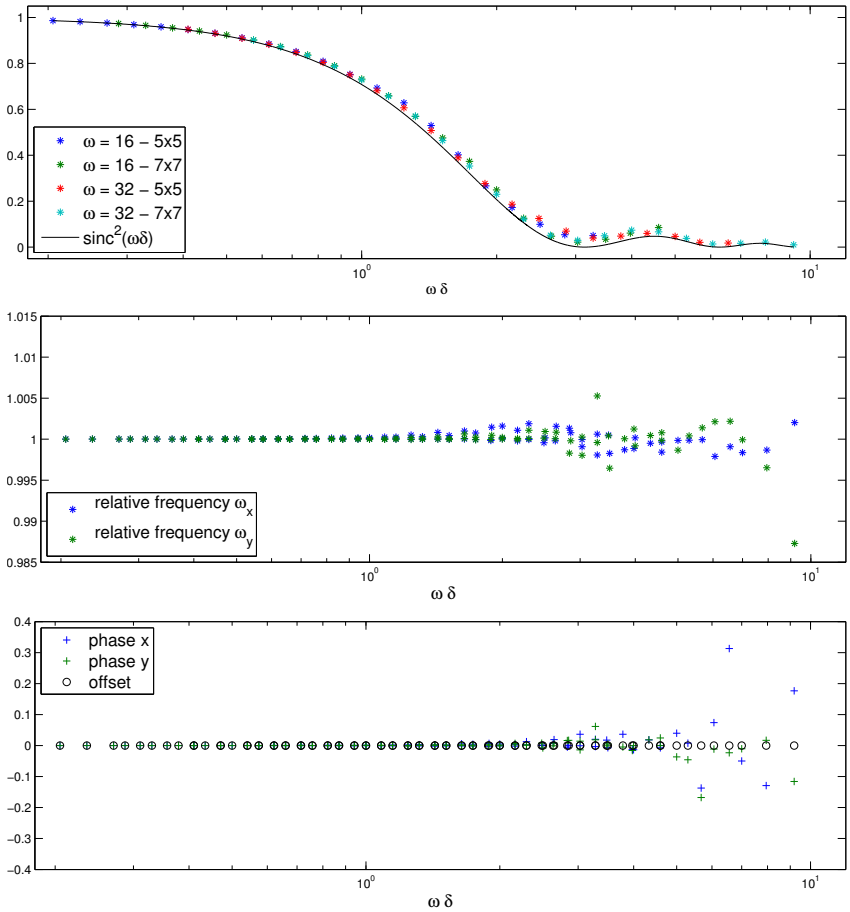


Figure 5.9: *Top:* Reconstructed amplitude as fraction of the true amplitude compared to theoretical prediction in 2D. *Middle:* Relative frequencies. *Bottom:* Relative phase and offset.

Figure 5.10: *Left:* Sample image of the lion head sculpture. *Right:* Low-resolution VRIP reconstruction.

5.4.3 Application to Real-World Example

After presenting all the theoretical results and experiments validating the results in practice, we want to exploit the new insights within a real-world application. In the following we enhance a single-scale multi-view stereo reconstruction. For that purpose we create a 3D model of a lion head sculpture using the following pipeline. We register 225 photographs [Snavely et al. 2008] of a lion head sculpture, reconstruct a depth map for a subsets of 41 views with image patch size of 7×7 pixels [Goesele et al. 2007], and merge the depth maps into a global model using VRIP [Curless and Levoy 1996] (see Figure 5.10). Hereby, we create two different versions, a low-resolution model using downscaled photos (halved image dimensions) for depth map reconstruction and a high-resolution model using full image resolution. We convert a cut-out of the models into a height field and smoothly interpolate to a constant value and zero gradient at the borders minimizing second order derivatives. This leads to a periodical signal which is the input to a 2D Fourier transform. For all frequencies, we compute the inverse MTF using our model and scale up the frequencies accordingly to invert the amplitude loss during reconstruction. Since our experiments showed significant noise and thus deviation from the ideal MTF for the real-world test target, we clamp the inverse MTF. We use $MTF_\delta(\omega)^{-1} = \min(0.6^{-1}, \mathrm{sinc}(\omega\delta)^{-1})$ (Figure 5.11). We also apply a smooth low-pass filter that suppresses high-frequencies where the patch size is smaller than the wavelength. Finally, the inverse Fourier transform is performed. Figures 5.12 and 5.13 show how details are emphasized through the inversion of our stereo transfer function. Difference images in Figure 5.14 show a quantitative comparison where some regions are

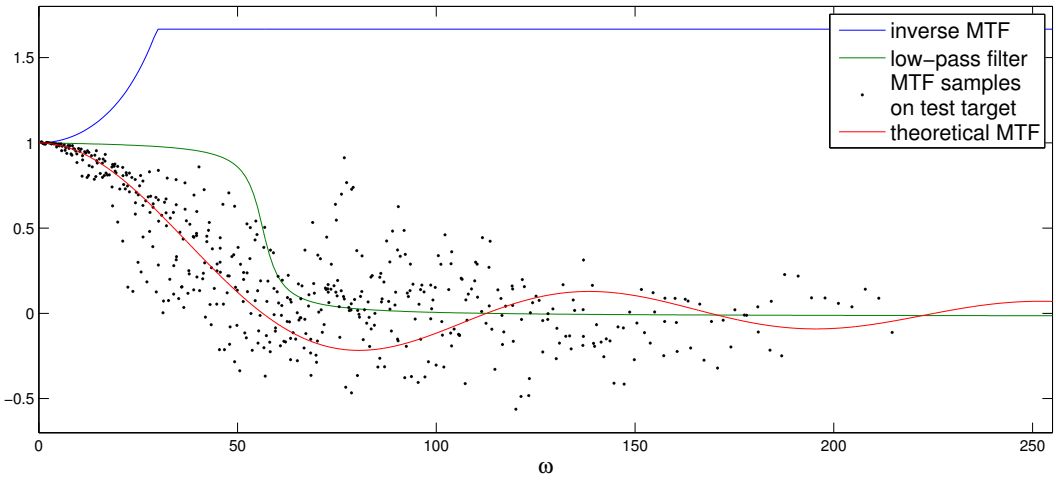
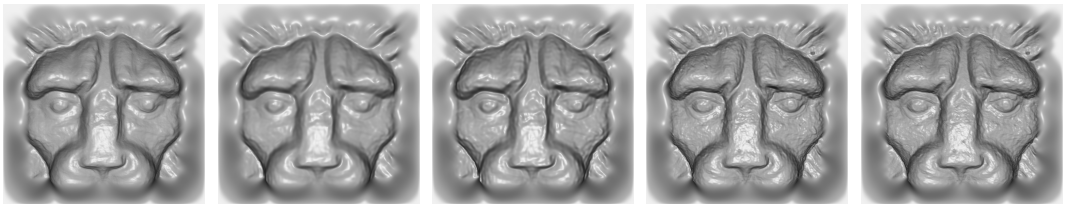


Figure 5.11: Slice of the 2D inverse MTF and of the low-pass filter used for the lion head experiment.

improved whereas others become worse.

5.5 Discussion

We introduce a theoretical model of patch-based stereo, modeling the reconstruction process as a linear system, and validate it on synthetic and real input using an existing multi-view stereo system. We demonstrate that there is a *significant amplitude loss* and even an inversion of amplitudes which has *not been modeled before* in any of the existing reconstruction pipelines. The real-world application example gives a first clue of how this could improve the reconstruction quality in a practical system. Inevitably, the experiments show some limitations. First, modeling the reconstruction process as finding the depth and orientation of a patch that minimizes the least squares distance to the true surface leaves out the complex interaction between the surface texture and the reconstruction. This may yield artifacts when the MTF is inverted. Second, the noise introduced in the reconstructions may of course limit the ability to invert the amplitude loss. Finally, practical applicability is limited because of the nature of the global Fourier method causing problems with depth discontinuities (occlusion), finite image size (periodicity assumption), and incomplete reconstructions. The lion head example is therefore only a starting point of how geometry can be reconstructed faithfully using our amplitude loss compensation.



a) *b)* *c)* *d)* *e)*

Figure 5.12: Results on a height field created from the lion head VRIP model. *a)* Low-resolution reconstruction. *b)* Removed high-frequency noise. *c)* Inverted amplitude loss up to a certain scale. *d)* Smoothed high-resolution reconstruction. *e)* High-resolution reconstruction.

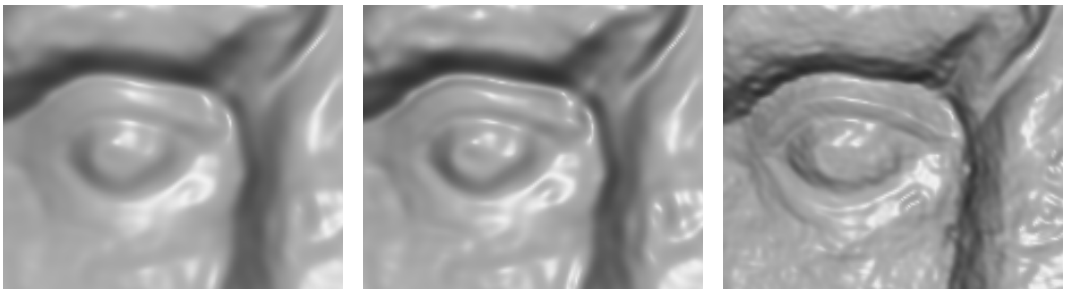


Figure 5.13: *From left to right:* Magnification of a region around the left eye in Figure 5.12 *b)*, *c)*, and *d)*, clearly showing how our proposed method improves the details, e.g., of the eyelid. See Figure 5.14 for a quantitative visualization of the differences.

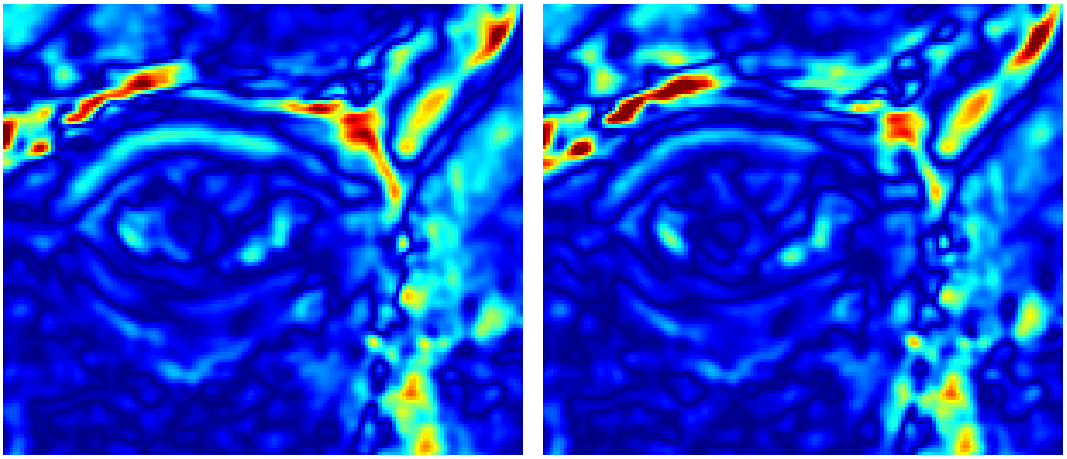


Figure 5.14: Absolute depth differences of results shown in Figure 5.13 left/middle compared to Figure 5.13 right. Note the changes around the eyelid and the nose.

6 Weighted Patch-based Reconstruction

Contents

SURFACE reconstruction using patch-based multi-view stereo commonly assumes that the underlying surface is locally planar. This is typically not true so that least squares fitting of a planar patch leads to systematic errors which are of particular importance for multi-scale surface reconstruction. In the previous chapter we determined the modulation transfer function of a classical patch-based stereo system. Our key insight was that the reconstructed surface is a box-filtered version of the original surface. Since the box filter is not a true low-pass filter this causes high-frequency artifacts. In this chapter, we propose an extended reconstruction model by weighting the least squares fit of the 3D patch. We show that if the weighting function meets specified criteria the reconstructed surface is the convolution of the original surface with that weighting function. A choice of particular interest is the Gaussian which is commonly used in image and signal processing but left unexploited by many multi-view stereo algorithms. Finally, we demonstrate the effects of our theoretic findings using experiments on synthetic and real-world data sets.

6.1 Introduction

The basis of virtually all multi-view stereo algorithms are correspondences found between images. Hereby, the de facto standard is to find a planar patch in 3D whose projected region in (some of) the images is photo-consistent, i.e., looks similar. There are many ways to measure photo-consistency including normalized cross-correlation (NCC) or the sum of squared differences (SSD, see Hu and Mordohai [Hu and Mordohai 2012] for an overview and evaluation of different measures). Whatever measurement used, the underlying assumption is that the original surface is locally planar or even has constant depth in the patch area. This leads to a systematic error in reconstruction which becomes especially important when combining multi-scale data [Bellocchio et al. 2013, Fuhrmann and Goesele 2011]. In the previous chapter we analyzed this systematic error and proposed a reconstruction model where the 3D patch is fitted to the original surface in a least squares sense. In the resulting linear system we identified the modulation transfer function to be a sinc. In other words, the reconstructed surface is equal to a convolution of the original surface with a box filter. Since this is no true low-pass filter it causes high-frequency artifacts such as amplitude inversion for some frequencies.

In this chapter, we develop an extended reconstruction model by weighting the fitting of the 3D patch. We derive constraints on the weighting function to ensure that the reconstructed surface is a convolution of the original surface with that weighting function. As a particular result, we will see that uniform weighting used in the previous chapter causes the box filter effect. A much better choice for the weighting function fulfilling the derived constraints and allowing for true low-pass filtered reconstructions is the Gaussian, which is widely used in the imaging domain. When using different patch sizes (e.g., due to different image resolution or camera-object distances) the reconstructions reflect different levels of the scale space representation of the true surface. We show for one popular multi-view stereo algorithm [Goesele et al. 2007] how to implement the weighting and discuss results on synthetic as well as real-world data sets. Our findings may influence a broad range of algorithms in multi-view stereo but also in the field of multi-scale surface reconstruction [Fuhrmann and Goesele 2011, Furukawa et al. 2010, Gargallo and Sturm 2005, Mücke et al. 2011] or geometry super-resolution [Goldluecke and Cremers 2009, Yang et al. 2007]. In chapter 7 we present a surface reconstruction framework that handles input data originating from a weighted multi-view stereo algorithm and exploit the results presented below.

In summary the contributions of this chapter are

- the generalization of a previously presented reconstruction model for (multi-view) stereo by introducing weights,
- the theoretical derivation of the (predicted) reconstructed surface without the detour in frequency space, and
- we show how a weighting, e.g., a Gaussian, can be implemented for a common multi-view stereo algorithm which expectably improves the frequency behavior of the reconstruction.

6.2 Related Work

While there is a large body of work on multi-view stereo (see, e.g., the survey paper and the constantly updated benchmark by Seitz at al. [Seitz et al. 2006, Seitz et al. 2013]), the study of multi-scale depth reconstruction has long been neglected. In the previous chapter we introduced a theoretical reconstruction model and determined the modulation transfer function of patch-based stereo systems. We also discussed the (loosely) related work on multi-scale analysis of (multi-view) stereo to which we refer the reader for a more extensive discussion. Our current work builds upon this reconstruction model and demonstrates how more freedom in the reconstruction outcome is possible. As one particular result, we demonstrate that multi-view stereo can yield a scale space representation of the underlying geometry. In contrast to the previous chapter, we now derive our results directly in geometry space without operation (at least in an intermediate step) in frequency space.

Our work is also related to existing work on patch-based photo-consistency measures. An overview and evaluation of confidence measures used in (multi-view) stereo is given by Hu and Mordohai [2012]. In all their cost computations, however, a square patch of $N \times N$ pixels is used and all pixels are weighted uniformly. If we assume all measures aim at fitting a patch in 3D space, they all result in a box filter. Kanade and Okutomi [1994] already tried to find optimal size and shape of the patch but still only used rectangular shapes. Habbecke and Kobbelt [2007] propose a multi-view stereo system where matching is performed on circular disks in object space. The size of the disks is selected to achieve a minimum intensity variance on each disk. Totally different shapes are achieved by Micusik and Koseka [Micusik and Kosecka 2008] whose approach is suited for man-made environments with many planar surfaces. Here, the reference view is first segmented into superpixels, that are assumed to be planar in object space, and matching is then performed using those superpixels. Thus the shape of the matching

window is adapted to the local scene structure and texture. Yoon and Kweon [2005] were probably the first to compute weights for each pixel in the patch that steer the influence of that pixel in the matching process. Their weights are dependent on the color similarity and the spatial distance from the center pixel. Hosni et al. [2009] improve on that by computing weights using the geodesic distance transform. In contrast to all these efforts, we investigate the influence of a specific weighting on the reconstructed geometry and derive the resulting (multi-scale) behavior of the resulting surface.

6.3 Theoretical Considerations

6.3.1 Extension of the Reconstruction Model

In this chapter, we build upon the reconstruction model introduced in Section 5.3 of the previous chapter. We describe the process of photometric consistency optimization between images (e.g. using normalized cross-correlation (NCC), or sum of squared differences (SSD)) as a geometric least squares fitting of a planar patch to the unknown geometry. To obtain the reconstruction at some point x, a line segment (parameterized by slope m and offset n) with extent 2δ is fitted to the geometry in a least squares sense minimizing the energy

$$E(m,n,x) = \int_{x-\delta}^{x+\delta} (mt+n-f(t))^2 dt. \tag{6.1}$$

The reconstructed surface is then represented by the central patch points. For this model we determined the modulation transfer function which turned out to be a sinc which is equivalent to a convolution with a box filter. In the following we will show that the reason for this result is the uniform weighting of pixels during optimization. We suggest the following extension of the reconstruction model: Instead of considering each point in $[x-\delta, x+\delta]$ uniformly we introduce a weighting function g allowing for different areas of influence. Consequently, we alter the energy function to

$$E(m,n,x) = \int_{-\infty}^{\infty} g(x-t)(mt+n-f(t))^2 \, dt \tag{6.2}$$

where $g(t)$ is a weighting function. Note that with $g(t) = \mathbf{1}_{[-\delta,\delta]}$ this is equal to the former energy in Eq. 6.1. This weighting function could be implemented as a weighting of the pixels during photo-consistency optimization. In Section 6.4 we will demonstrate this using a specific multi-view stereo algorithm. In the following subsection, we derive theoretically how this weighting function affects the reconstructed surface.

6.3.2 Reconstruction in 2D

For the sake of simplicity, we first look at a surface in 2D (a line) as illustrated in Figure 5.2. For now, we put no further constraints on $g(t)$ except for integrability. Later on, we will discuss further desirable properties. Minimizing E in Equation 6.2 requires taking the partial derivatives with respect to m and n:

$$\begin{aligned}\partial_m E &= 2\int_{-\infty}^{\infty} g(x-t)t(mt+n-f(t))\,dt \qquad (6.3)\\ &= 2m\int_{-\infty}^{\infty} g(x-t)t^2\,dt + 2n\int_{-\infty}^{\infty} g(x-t)t\,dt - 2\int_{-\infty}^{\infty} g(x-t)tf(t)\,dt\\ \partial_n E &= 2\int_{-\infty}^{\infty} g(x-t)(mt+n-f(t))\,dt \qquad (6.4)\\ &= 2m\int_{-\infty}^{\infty} g(x-t)t\,dt + 2n\int_{-\infty}^{\infty} g(x-t)\,dt - 2\int_{-\infty}^{\infty} g(x-t)f(t)\,dt\end{aligned}$$

We introduce a short notation for the zeroth, first and second moment of g

$$\mu_0 = \int_{-\infty}^{\infty} g(t)\,dt \qquad \mu_1(x) = \int_{-\infty}^{\infty} g(x-t)t\,dt \qquad \mu_2(x) = \int_{-\infty}^{\infty} g(x-t)t^2\,dt \quad (6.5)$$

and abbreviate the other convolution integrals using

$$(g*\cdot f)(x) = \int_{-\infty}^{\infty} g(x-t)tf(t)\,dt \qquad (6.6)$$

$$(g*f)(x) = \int_{-\infty}^{\infty} g(x-t)f(t)\,dt. \qquad (6.7)$$

W.l.o.g. we can assume that $\mu_0 = 1$ which corresponds to normalizing the weighting function g. Under the condition that $\mu_2(x) \neq 0$ we set the partial derivatives to zero and transpose the equations:

$$m = \frac{(g*\cdot f)(x) - n\mu_1(x)}{\mu_2(x)} \qquad (6.8)$$

$$n = (g*f)(x) - m\mu_1(x) \qquad (6.9)$$

We can now solve for m and n which leads to

$$\begin{aligned} m &= \frac{(g*\cdot f)(x) - ((g*f)(x) - m\mu_1(x))\mu_1(x)}{\mu_2(x)} \\ \Leftrightarrow\ m &= \left(1 - \frac{\mu_1(x)^2}{\mu_2(x)}\right)^{-1} \left(\frac{(g*\cdot f)(x)}{\mu_2(x)} - \frac{(g*f)(x)\mu_1(x)}{\mu_2(x)}\right) \\ &= \frac{(g*\cdot f)(x) - (g*f)(x)\mu_1(x)}{\mu_2(x) - \mu_1(x)^2} \qquad (6.10) \\ n &= (g*f)(x) - \frac{(g*\cdot f)(x) - (g*f)(x)\mu_1(x)}{\mu_2(x) - \mu_1(x)^2}\mu_1(x) \\ &= \frac{(g*f)(x)\mu_2(x) - (g*\cdot f)(x)\mu_1(x)}{\mu_2(x) - \mu_1(x)^2} \qquad (6.11) \end{aligned}$$

Since the final surface is represented by the central patch points it can be written as

$$mx + n = \frac{(g*\cdot f)(x)(x - \mu_1(x)) + (g*f)(x)(\mu_2(x) - x\mu_1(x))}{\mu_2(x) - \mu_1(x)^2}. \qquad (6.12)$$

Though valid for very general weighting functions g this result is not very satisfactory. On closer inspection we see that when $\mu_1(x) = x$, which is true for all normalized symmetric functions g, it can be easily simplified to

$$mx + n = (g*f)(x). \qquad (6.13)$$

In other words, every function g with $\mu_0 = 1$, $\mu_1(x) = x$, $\mu_2(x) \neq 0$, and $\mu_2(x) \neq x^2$, used to weight the least squares fitting results in a reconstruction that is the convolution of the true surface with g. Note, that a uniform weighting naturally leads to the convolution with a box filter (cf. Chapter 5) in this framework.

6.3.3 Building a Scale Space Representation

The derived constraints for the weighting function obviously allow for many different choices. One of particular interest is the Gaussian since convolutions with Gaussians are well studied and widely applied, e.g., in the image domain. If we set g to be a normalized Gaussian with standard deviation σ

$$g(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-t^2}{2\sigma^2}\right). \qquad (6.14)$$

we obtain the following moments

$$\mu_0 = 1 \qquad \mu_1(x) = x \qquad \mu_2(x) = \sigma^2 + x^2. \qquad (6.15)$$

That is, the normalized Gaussian fulfills our constraints and we can determine the slope m and offset n of the fitted patch at each point x by

$$m = \frac{(g * \cdot f)(x) - (g * f)(x)x}{\sigma^2} \tag{6.16}$$

$$n = \frac{(g * f)(x)(\sigma^2 + x^2) - (g * \cdot f)(x)x}{\sigma^2}. \tag{6.17}$$

In order to create a scale space representation of the underlying surface we need to use Gaussians with varying standard deviations σ. However, during reconstruction we can influence σ only to a limited extent because it depends on the scene depth, image resolution and focal length of the camera. In that sense, if we reconstruct depth maps of the same geometry using a variety of images results in a natural variation of the standard deviation σ in real-world space. The only parameter one can actively steer is the standard deviation σ_i (linked with the window size due to approximation and clamping of the Gaussian) in image space used for patch-based optimization. When selecting σ_i one often has a rough depth estimate and also the camera parameters are known from registration. With that it is possible to indirectly steer the standard deviation σ in world space at least to a limited extent, e.g., for parts of the scene with different depths. In Section 6.4 we will conduct some experiments with varying the standard deviation σ_i but we first transfer our results into 3D.

6.3.4 Reconstruction in 3D

For the reconstruction in 3D we assume the 2D geometry is described as a height field $z = f(x, y)$. To obtain the reconstruction at some point (x, y), we fit a patch (surface segment) that is parameterized by 2 slopes m_1 and m_2 and an offset n. Again, the weighting function g allows for different areas of influence. As a result we now have the following energy

$$E(m_1, m_2, n, x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x - t, y - s)(m_1 t + m_2 s + n - f(t, s))^2 \, dt \, ds. \tag{6.18}$$

Minimizing E requires taking the partial derivatives with respect to m_1, m_2, and n:

$$\partial_{m_1} E = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 2t g(x - t, y - s)(m_1 t + m_2 s + n - f(t, s)) \, dt \, ds \stackrel{!}{=} 0 \tag{6.19}$$

$$\partial_{m_2} E = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 2s g(x - t, y - s)(m_1 t + m_2 s + n - f(t, s)) \, dt \, ds \stackrel{!}{=} 0 \tag{6.20}$$

$$\partial_n E = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 2 g(x - t, y - s)(m_1 t + m_2 s + n - f(t, s)) \, dt \, ds \stackrel{!}{=} 0 \tag{6.21}$$

Similar to the reconstruction in 2D, we introduce the short notation μ_{00}, μ_{10}, μ_{01}, μ_{20}, μ_{11}, and μ_{02} for the moments of g with respect to x and y, respectively.

$$\mu_{00} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(t,s)\,dt\,ds, \qquad \mu_{10} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x-t,y-s)t\,dt\,ds \tag{6.22}$$

$$\mu_{01} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x-t,y-s)s\,dt\,ds, \quad \mu_{20} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x-t,y-s)t^2\,dt\,ds \tag{6.23}$$

$$\mu_{11} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x-t,y-s)st\,dt\,ds, \mu_{02} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x-t,y-s)s^2\,dt\,ds \tag{6.24}$$

For the sake of clarity we chose an even shorter abbreviation for the other convolution integrals:

$$\mathrm{gtf} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} tg(x-t,y-s)f(t,s)\,dt\,ds \tag{6.25}$$

$$\mathrm{gsf} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} sg(x-t,y-s)f(t,s)\,dt\,ds \tag{6.26}$$

$$\mathrm{gf} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x-t,y-s)f(t,s)\,dt\,ds. \tag{6.27}$$

Again, we can normalize g such that $\mu_{00} = 1$. With this notation we can rewrite Eqs. (6.19)-(6.21) as

$$\partial_{m_1} E = 2(m_1\mu_{20} + m_2\mu_{11} + n\mu_{10} - \mathrm{gtf}) \stackrel{!}{=} 0 \tag{6.28}$$

$$\partial_{m_2} E = 2(m_1\mu_{11} + m_2\mu_{02} + n\mu_{01} - \mathrm{gsf}) \stackrel{!}{=} 0 \tag{6.29}$$

$$\partial_n E = 2(m_1\mu_{10} + m_2\mu_{01} + n - \mathrm{gf}) \stackrel{!}{=} 0 \tag{6.30}$$

Solving these equations for m_1, m_2, and n yields

$$\alpha m_1 = \mathrm{gf}(\mu_{02}\mu_{10} - \mu_{01}\mu_{11}) + \mathrm{gsf}(\mu_{11} - \mu_{01}\mu_{10}) + \mathrm{gtf}\left(\mu_{01}^2 - \mu_{02}\right) \tag{6.31}$$

$$\alpha m_2 = \mathrm{gf}(\mu_{01}\mu_{20} - \mu_{10}\mu_{11}) + \mathrm{gsf}\left(\mu_{10}^2 - \mu_{20}\right) + \mathrm{gtf}(\mu_{11} - \mu_{01}\mu_{10}) \tag{6.32}$$

$$\alpha n = \mathrm{gf}\left(\mu_{11}^2 - \mu_{02}\mu_{20}\right) + \mathrm{gsf}(\mu_{01}\mu_{20} - \mu_{10}\mu_{11}) + \mathrm{gtf}(\mu_{02}\mu_{10} - \mu_{01}\mu_{11}) \tag{6.33}$$

where $\alpha = \mu_{20}\mu_{01}^2 - 2\mu_{10}\mu_{11}\mu_{01} + \mu_{02}\mu_{10}^2 + \mu_{11}^2 - \mu_{02}\mu_{20}$. Plugging in these expressions in the patch $P = m_1 x + m_2 y + n$, we obtain

$$P = \frac{1}{\alpha}\big(\mathrm{gf}\big(\mu_{11}^2 - \mu_{02}\mu_{20} - \mu_{01}\mu_{11}x + \mu_{02}\mu_{10}x - \mu_{10}\mu_{11}y + \mu_{01}\mu_{20}y\big) + \tag{6.34}$$

$$\mathrm{gsf}\big(-\mu_{11}\mu_{10} + \mu_{01}\mu_{20} - \mu_{01}\mu_{10}x + \mu_{11}x + \mu_{10}^2 y - \mu_{20}y\big) + \tag{6.35}$$

$$\mathrm{gtf}\big(-\mu_{11}\mu_1 + \mu_{02}\mu_{10} + \mu_{01}^2 x - \mu_{02}x - \mu_{10}\mu_{01}y + \mu_{11}y\big)\big). \tag{6.36}$$

Taking symmetric filters yields $\mu_{10} = x$ and $\mu_{01} = y$. Then immediately one gets

$$P = \mathrm{gf} \tag{6.37}$$

Of course we can use a classical anisotropic Gaussian characterized by σ and τ

$$g(t,s) = \frac{1}{2\pi\sigma\tau} \exp\left(\frac{-t^2}{2\tau^2} + \frac{-s^2}{2\sigma^2}\right) \tag{6.38}$$

because the moments are $\mu_{00} = 1$, $\mu_{10} = x$, $\mu_{01} = y$, $\mu_{02} = x^2 + \tau^2$, $\mu_{11} = xy$, $\mu_{02} = y^2 + \sigma^2$.

6.4 Experiments

In order to verify our theoretic findings in practice we now conduct some experiments. We hereby chose the depth map reconstruction method of Goesele et al. [2007] because it does a pure photo-consistency optimization (going back to Gruen and Baltsavias [1988]) to find depth and normal for a certain pixel and has no regularization force. For a small region around a pixel i, j in a reference view I_R the method aims to find depth d and normal $\vec{n}$ of the associated 3D patch such that it is photo-consistent with a set of neighboring views I_k. The algorithm minimizes (see [Goesele et al. 2007, Sec. 6.2] ignoring the color scale)

$$\sum_{k,i,j} [I_R(s+i, t+j) - I_k(P_k^{d,\vec{n}}(s+i, t+j))]^2 \tag{6.39}$$

where P_k describes the projection of a pixel from the reference view in the neighbor view I_k according to some depth d and normal $\vec{n}$. We implement the weighting on the least squares patch fit by weighting the pixels, i.e., we compute a weighted SSD:

$$\sum_{k,i,j} g(i,j)[I_R(s+i, t+j) - I_k(P_k^{d,\vec{n}}(s+i, t+j))]^2. \tag{6.40}$$

The remaining question is whether this weighted photo-consistency optimization still reflects the process of weighted least squares fitting as described by Eq. 6.2. We test this using a synthetic data set because of two reasons: First, we can assure that our results are not affected by registration errors but solely reflect the photometric consistency optimization, and second, we know the ground truth surface and are able to compute the predicted reconstruction according to our model. Our ground truth surface is created as a random sum of one-dimensional B-Splines extruded into the third dimension. We

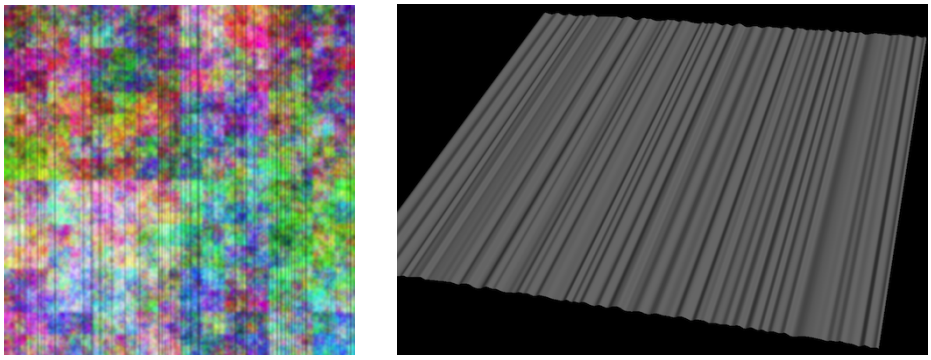


Figure 6.1: Left: The central view of our synthetic data set. Right: The underlying mesh (shaded) used to render the views.

then render five different views (one central view looking perpendicular onto the surface and four views distributed uniformly around it with a parallax of 35°) of this scene using the PBRT system [Pharr and Humphreys 2012] while a random texture is mapped onto the surface to guarantee matching success at all pixels (see Fig. 6.1). For the central view we now reconstruct a depth map by using the other four views as neighbors and minimizing the weighted SSD from Eq. 6.40. We start the optimization for each pixel with the depth value obtained from PBRT and the normal representing a fronto-parallel patch. To reduce noise we average the reconstructed values along the constant dimension. Fig. 6.2 shows the reconstructions using a uniform weighting function. The quadratic windows in image space are 11 (blue), 21 (green), 41 (red), and 61 (cyan) pixels wide which corresponds to a patch size (2δ) of 0.06,0.12,0.24, and 0.36 in world coordinates, respectively. We also plotted the predicted reconstructions, i.e., convolutions of the original surface with box filters of the corresponding width. Overall, the reconstruction is close to the prediction although there is some local deviation. The best conformity is achieved for the small patch size which can also be seen in Table 6.1 where we computed the mean deviation. Note the occasional amplitude inversion visible in the prediction as well as the reconstruction, in particular for the largest filter at around -1.4.

In Fig. 6.3 we used Gaussian weighting with increasing standard deviation which leads to a scale space representation of the underlying surface. The window sizes are the same used for the uniform weighting and we always chose the standard deviation σ such that $\delta = 2.5\sigma$. That is, in world coordinates we used $\sigma = 0.012, 0.024, 0.048, 0.072$. We can see from the figure and also by studying the numbers in Table 6.1 that the deviation from the prediction again increases for larger σ.

Finally, we show reconstruction results on real world data. Figure 6.4 (top left) shows

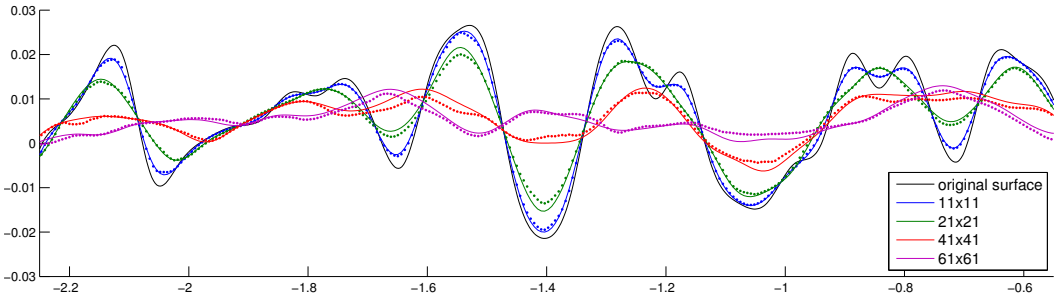


Figure 6.2: Multi-view stereo reconstruction using a uniform weighting with increasing patch size. The black line denotes the original surface. The colored solid lines are the computed predictions while the corresponding dots are the reconstructed values.

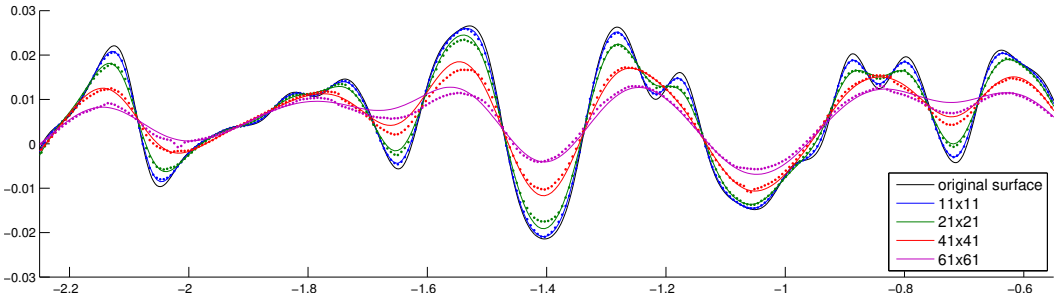


Figure 6.3: Reconstructing a scale space representation using a Gaussian weighting with increasing standard deviation (see text). The black line denotes the original surface. The colored solid lines are the computed predictions while the according dots are the reconstructed values.

an input image of the Notre Dame data set consisting of 715 images downloaded from the Internet. We use Snavely et al. [2008] to register them and compute depth maps for the shown image using different weightings and window sizes. The middle and bottom row show reconstructions obtained using uniform and Gaussian weighting, respectively. Although hard to jugde, the Gaussian weighting seems to produce slightly more noise and less complete reconstructions. On the other hand it better preserves the low frequencies. One must consider though, that the algorithm [Goesele et al. 2007] was tuned to work well with the uniform weighting and on a broad range of data sets. That is, playing with the parameters in the optimization or view selection might result in more favorable results for the Gaussian weighting.

Patch size in pixels	mean deviation (L_1-norm)	
	uniform weighting	Gaussian weighting
11×11	$1.9 \cdot 10^{-4}$	$1.3 \cdot 10^{-4}$
21×21	$4.1 \cdot 10^{-4}$	$2.8 \cdot 10^{-4}$
41×41	$6.9 \cdot 10^{-4}$	$5.8 \cdot 10^{-4}$
61×61	$6.3 \cdot 10^{-4}$	$7.0 \cdot 10^{-4}$

Table 6.1: Mean deviation of the reconstruction from the theoretical predicted surface (see Figs. 6.2&6.3).

6.5 Discussion

This chapter extends a recently introduced model for patch-based depth reconstruction by adding a weighting function. We derive criteria on the weighting function such that we can predict the reconstructed surface as the convolution of the true surface with the applied weighting function. This includes using a Gaussian instead of a uniform weighting during reconstruction which corresponds to a Gaussian instead of a box filter in geometry space. In contrast to previous methods, we achieve a true low-pass filter avoiding the introduction of systematic high-frequency artifacts. Future work definitely includes to further investigate the correlation between weighted photo-consistency optimization and weighted least squares fitting of a planar patch to the geometry.

Our findings are applicable in a broad range of applications. In contrast to the results of the previous chapter, we now give a local characterization of the reconstruction outcome at the same time offering more flexibility caused by the weighting. Multi-scale surface reconstruction methods such as [Fuhrmann and Goesele 2011, Furukawa et al. 2010, Gargallo and Sturm 2005, Mücke et al. 2011] could take that knowledge into account when combining data from multiple depth maps. But also geometry super-resolution methods [Goldluecke and Cremers 2009, Yang et al. 2007] can benefit from our findings. Since we provide evidence for a generative model it is now possible to adapt well established methods from imaging, e.g., Bayesian super-resolution [Pickup et al. 2007], to the geometry reconstruction context.

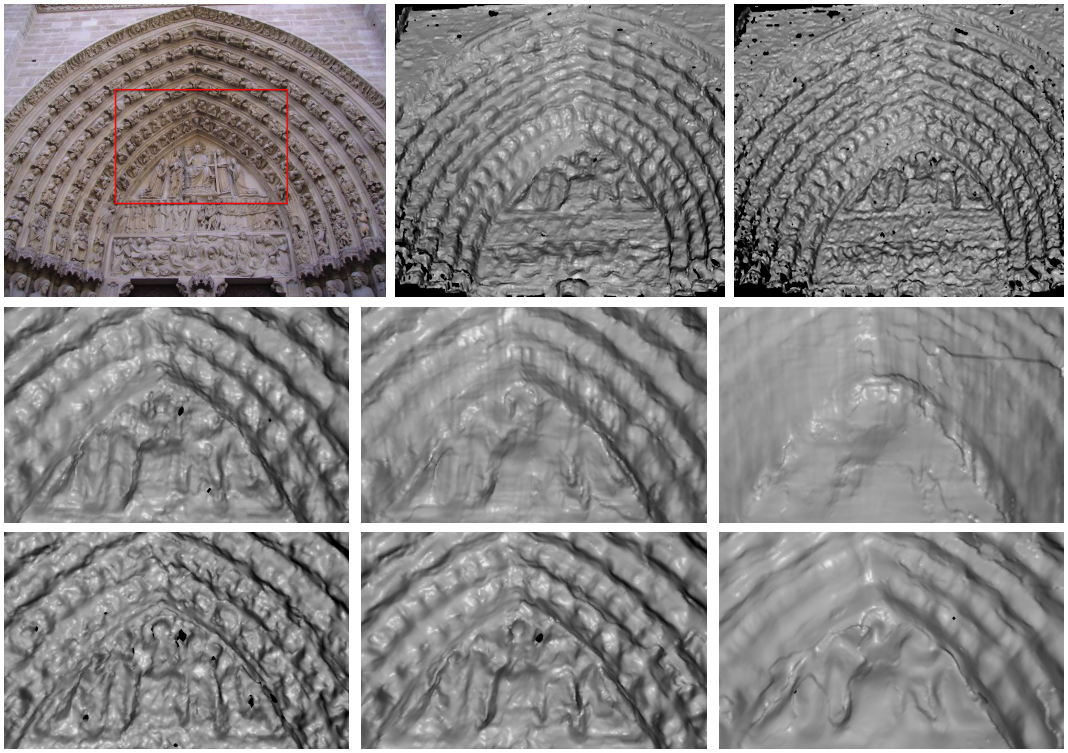


Figure 6.4: *Top left*: Input image of the Notre Dame data set. The red box is roughly the area seen in the bottom rows. *Top middle,right*: Full rendered view of reconstructed depth map using uniform (middle) and Gaussian weighting (right) and a window size in images space of 7×7 pixels. *Middle+Bottom*: Enlarged area roughly corresponding to red box (top left) of the reconstructed depth map. We applied uniform (middle) and Gaussian weighting (bottom) using window sizes of 7×7, 11×11, and 21×21 pixels (from left to right) for reconstruction where the standard deviation of the Gaussian in image space is $\sigma_i = 1.2, 2.0, 4.0$.

7 Wavelet-based Surface Reconstruction

Contents

Multi-view stereo reconstruction techniques yield inherently multi-scale point data typically fed into surface reconstruction algorithms. Following the intuition of scale space we assume that sample points originate from smoothed versions of the original surface. The smoothing can be characterized by a smoothing kernel that suppresses fine-scale structures. In this chapter, we propose a surface reconstruction framework that correctly handles this multi-scale input data. We represent the surface using a multi-resolution analysis allowing us to reconstruct scales separately and to merge the sample points in frequency space. With an underlying wavelet basis we are able to locally model surface detail according to the surface properties or sample distribution. We first demonstrate the effectiveness of our method on a synthetic data set with known smoothing. For real-world data obtained by multi-view stereo we estimate the smoothing kernel and present reconstruction results with enhanced detail.

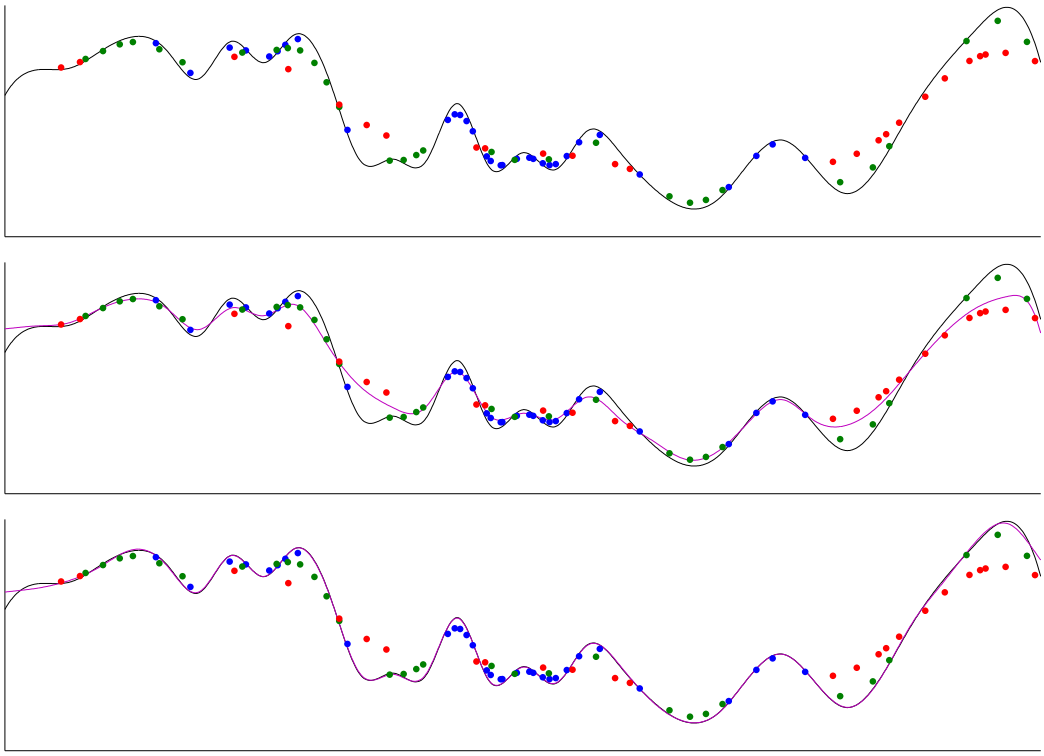


Figure 7.1: True surface (black) and multi-scale sample points (red–coarse, green–medium, blue–fine). *Top:* Input data. *Middle:* Reconstruction (magenta) treating all sample points equally. *Bottom:* Our reconstruction which takes scale into account and follows the true surface more clearly.

7.1 Introduction

Surface reconstruction from (unorganized) sample points is a well-researched area but also a continuous challenge. Popular methods include the pioneering work of Hoppe *et al*. [1992], range image integration (VRIP) proposed by Curless and Levoy [1996], and Poisson surface reconstruction by Kazhdan *et al*. [2006]. Recent papers [Fuhrmann and Goesele 2011, Manson et al. 2008, Shalom et al. 2010] give a detailed overview of the various methods available today. The focus of this chapter lies on the multi-scale component inherent to many reconstruction techniques such as multi-view stereo. These approaches are able to deal with large scenes, for example comprising entire cities [Agarwal et al. 2009], and a mixture of various cameras ranging from mobile phones to digital SLRs. Drastically different object-to-camera distances and varying image resolutions automatically yield multi-scale sample points. When talking about scales of a surface we typically think of gradually removing detail structures of the original surface with a

low-pass filter, which we model using a smoothing kernel. The main characteristic of multi-scale input data is that the samples are taken from successively smoothed versions (i.e., scales) in contrast to the simple case where all samples originate from the same scale (see the reconstruction in Fig. 7.1 top). In fact, it is commonly assumed that the input points are real point samples of the original surface implying that no or very little smoothing is involved (Fig. 7.1 middle). The first, and to our knowledge the only, to consider the multi-scale properties of sample points in a surface reconstruction algorithm are Fuhrmann and Goesele [2011]. They essentially remove coarse-scale data points (originating from strongly smoothed versions of the original surface) in areas where fine-scale points (less smoothed) with high confidence are available. Using this heuristic they are able to achieve impressive results on real world data sets. However, they rely on the correlation of resolution and scale suggesting that fine-scale sample points are usually present in higher resolution than coarse-scale samples. Also, discarding samples is a binary decision and information might be thrown away that could have been useful to close holes or even improve the fine-scale reconstruction. In summary, the fundamental problem of how to correctly merge multi-scale data points, i.e., combine the coarse- and fine-scale data instead of discarding the former, is still not convincingly solved.

In this chapter we propose a reconstruction framework for 2.5D height field representations (Sec. 7.3) that explicitly models and incorporates the multi-scale properties of the input data (Fig. 7.1 bottom). We use the concept of multi-resolution analysis (multi-scale approximation) of the original surface. With the generating scaling functions and wavelets we are able to simultaneously decompose the surface in space and frequency domain. Given sample points with known or approximated smoothing kernel we show how the original surface can be recovered correctly. Hereby, our surface representation allows for locally varying degree of detail according to surfaces shape and sample point distribution. For practical application (Sec. 7.4) we add a regularization term to the surface recovery and formulate an optimization problem. We further propose a specific wavelet representation and discuss the scale estimation in the context of multi-view stereo. Finally, we show results demonstrating the effectiveness of our method (Sec. 7.5) and conclude the chapter with an outlook on future work (Sec. 7.6).

7.2 Related Work

Classic surface reconstruction methods work on regularly sampled, some also on multi-resolution data points [Hoppe et al. 1992, Curless and Levoy 1996, Kazhdan et al. 2006,

Shalom et al. 2010]. The data is assumed to be single-scale which means that all points share the same noise model with the true surface as mean. A few recent approaches deviate from this paradigm. Mücke *et al*. [2011] use a Gaussian noise model but assign to each sample point a different standard deviation. They build a confidence volume represented in an octree and compute a minimum cut to reconstruct the surface (similar to other graph cut based methods [Boykov and Kolmogorov 2003, Hornung and Kobbelt 2006b, Sinha et al. 2007]). Fuhrmann and Goesele [2011] integrate depth maps, similar to VRIP [Curless and Levoy 1996], into a hierarchical signed distance field (hSDF). They subsequently prune the hSDF removing coarse-scale data in regions where fine-scale data is available. The final surface is then extracted using a variant of the marching tetrahedra algorithm. Bailer *et al*. [2012] handle the scale problem in a similar manner and also select locally the highest scale reconstruction available. Zach *et al*. [2007] integrate range images into a global signed distance field and add a regularization term that minimizes the total variation (L1-regularization) of the SDF. Some of these methods support multi-resolution representations with locally varying level-of-detail and are capable of producing impressive results even on uncontrolled multi-view stereo data sets. However, none of them combines data from different scales while modeling the different degree of smoothing.

Pauly *et al*. [2006] clarify the difference between multi-scale and multi-resolution surface representation. They use approximate low-pass filters to create a point-based multi-scale surface representation for the context of surface editing. Kazhdan [2005] incorporates Fourier theory for surface reconstruction. The method aims at recovering the characteristic function of the solid by reconstructing its Fourier coefficients. While theoretically well founded the method requires summing over all input points to compute each single Fourier coefficient. This is computationally extremely expensive and implies that a single point influences the entire model which is counterintuitive. It also requires some heuristics to process non-uniformly sampled data. In a recent paper, Digne *et al*. [2011] propose a scale space meshing method that implements the mean curvature motion (MCM) on the raw point set. They reconstruct a smooth mesh first and then revert the MCM. It would be interesting to investigate handling of multi-scale data with this approach.

Several authors proposed surface reconstruction methods using smooth basis functions possibly integrated in a wavelet space. In the early work of Pastor and Rodríguez [1999] spherical wavelets are used which naturally limits the application to objects that are topologically equivalent to a sphere. Carr *et al*. [2001] reconstruct smooth surfaces on the basis of smooth radial basis functions from noisy data. By computing the Fourier co-

efficients Kazhdan [2005] actually represents the indicator function using dilations and translations of the sine function. Manson *et al*. [2008] improve on this idea and apply wavelets instead, exploiting the local support to decrease complexity. A direct surface representation in Monge's form, as used in this chapter, was proposed by Johnson *et al*. [2009]. They use B-Splines and associated wavelets for scattered data reconstruction and give a theoretical error analysis. For better preserving depth discontinuities Ji *et al*. [2010] seek for a piecewise smooth approximation in tight wavelet frames. All of these and other related methods in scattered data interpolation do not tackle the problem of multi-scale input data as we do in this chapter. Also, the multi-scale structure of the basis functions is not exploited in order to adjust the granularity of the final reconstruction according to the input data.

7.3 Reconstruction Framework

The basis of our reconstruction framework is a surface representation that allows us to operate on different scales of the surface. With that we can model surfaces with locally varying detail, either due to the surface itself or due to the distribution of the sample points. The classic Fourier transform is unsuited due to boundary handling issues and the missing locality. The latter also implies a constant frequency resolution over the entire space without taking into account the actual sample distribution. This involves the risk to hallucinate high frequency details in regions that are not sampled at all. In the following we first introduce our surface representation and describe afterwards how the surface can be recovered correctly from multi-scale sample points.

7.3.1 Surface Representation

We use an explicit surface representation assuming that the surface can be parameterized as a height field $f(x)$. For simplicity the following derivation is for the 1D case $x \in \mathbb{R}$ but it can be easily extended to higher dimensions by applying the standard multi-dimensional wavelet construction described by Mallat [2008, Ch. 7.7]. We embed the surface in a multi-resolution analysis, written according to the notation of Stollnitz *et al*. [1996, Ch. 7] as

$$\mathbf{V}_0 \subset \mathbf{V}_1 \subset \mathbf{V}_2 \ldots \subset L^2(\mathbb{R}) \tag{7.1}$$

where $\mathbf{V}_0$ can be thought of containing very smooth surfaces and with increasing index j in V_j more detail can be added (see Fig. 7.2). Eventually all possible surfaces $f \in L^2(\mathbb{R})$

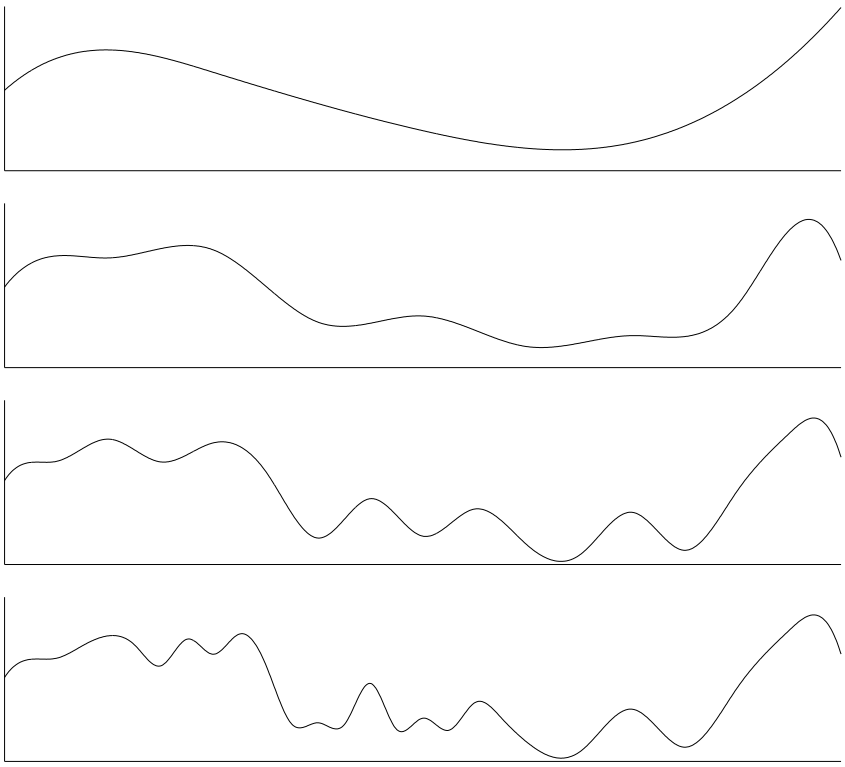


Figure 7.2: Multi-resolution analysis of a 1D surface. The detail level j increases from top to bottom and local surface details become visible.

are included. The complements of $\mathbf{V}_j$ in $\mathbf{V}_{j+1}$ are denoted by $\mathbf{W}_j$ such that

$$\mathbf{V}_{j+1} = \mathbf{V}_j + \mathbf{W}_j, j > 0. \tag{7.2}$$

The $\mathbf{V}_j$ are spanned by shifted and dilated versions $\phi_{j,l} = \phi(2^j x - l)$ of the father wavelet (or scaling function) ϕ and the $\mathbf{W}_j$ by shifted and dilated versions $\psi_{j,l}$ of the mother wavelet ψ, respectively. With that the surface f can be represented by its wavelet decomposition

$$f(x) = \sum_l c_{0,l}\phi_{0,l}(x) + \sum_{j=0}^{\infty}\sum_l d_{j,l}\psi_{j,l}(x) \tag{7.3}$$

where the $c_{0,l}$ denote the scaling function and the $d_{j,l}$ the wavelet coefficients. One can think of modeling the rough shape through the $c_{0,l}$ and then adding more and more details with increasing j by activating the $d_{j,l}$. Typically, the (effective) support of the $\psi_{j,l}$ decreases with increasing j so that surface details can be modeled locally. Since $\mathbf{V}_j = \mathbf{V}_0 + \mathbf{W}_0 + \ldots + \mathbf{W}_{j-1}$ one could also start with scaling functions of higher level. Also,

in practice one has to cut off somewhere resulting in the more general representation:

$$f(x) \approx \sum_l c_{j_0,l}\phi_{j_0,l}(x) + \sum_{j=j_0}^{j_{\max}} \sum_l d_{j,l}\psi_{j,l}(x). \tag{7.4}$$

Without loss of generality we will in the following assume $j_0 = 0$ and for convenience we will use the equal sign although we refer to the approximation.

7.3.2 Surface Recovery from Samples

Given ideal point samples $(x_i, y_i)_{i=1,\dots,N}$ from the surface with $y_i = f(x_i)$ we have a linear system of equations

$$y_i = \sum_l c_{0,l}\phi_{0,l}(x_i) + \sum_{j=0}^{j_{\max}} \sum_l d_{j,l}\psi_{j,l}(x_i) \tag{7.5}$$

and the coefficients $c_{0,l}, d_{j,l}, 0 \le j \le j_{\max}$ as unknown variables. We can rewrite Eq. (7.5) in matrix form

$$\begin{pmatrix} \phi_{0,l}(x_1) & \psi_{0,l}(x_1) & \dots & \psi_{j_{\max},l}(x_1) \\ \vdots & & & \vdots \\ \phi_{0,l}(x_N) & \psi_{0,l}(x_N) & \dots & \psi_{j_{\max},l}(x_N) \end{pmatrix} \begin{bmatrix} c_{0,l} \\ d_{0,l} \\ \vdots \\ d_{j_{\max},l} \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}. \tag{7.6}$$

When we introduce multi-scale samples, i.e., sample points from the gradually smoothed surface, we assume that for each sample (x_i, y_i) the convolution kernel g_i is known such that

$$y_i = (g_i * f)(x_i). \tag{7.7}$$

This is a very general setup since we do not commit ourselves to a particular smoothing kernel. In standard scale-space, with a Gaussian convolution, it is just the standard deviation σ_i that varies among the samples but here we allow for other kernels (e.g., Laplacians, splines, or box filters) as well. Note that ideal point samples are also covered by simply using the Dirac delta function $g_i(t) = \delta(t - x_i)$. With Eq. (7.7) the linear system changes to

$$\begin{aligned} y_i &= (g_i * f)(x_i) && (7.8) \\ &= \Big(g * \Big(\sum_l c_{0,l}\phi_{0,l} + \sum_{j=0}^{j_{\max}} \sum_l d_{j,l}\psi_{j,l} \Big) \Big)(x_i) \\ &= \sum_l c_{0,l}(g_i * \phi_{0,l})(x_i) + \sum_{j=0}^{j_{\max}} \sum_l d_{j,l}(g_i * \psi_{j,l})(x_i). && (7.9) \end{aligned}$$

Again, we write Eq. (7.9) in matrix form

$$\begin{pmatrix} \hat{\phi}^1_{0,l} & \hat{\psi}^1_{0,l} & \cdots & \hat{\psi}^1_{j_{\max},l} \\ \vdots & & & \vdots \\ \hat{\phi}^N_{0,l} & \hat{\psi}^N_{0,l} & \cdots & \hat{\psi}^N_{j_{\max},l} \end{pmatrix} \begin{bmatrix} c_{0,l} \\ d_{0,l} \\ \vdots \\ d_{j_{\max},l} \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \tag{7.10}$$

with $\hat{\phi}^i_{0,l} = (g_i * \phi_{0,l})(x_i)$ and $\hat{\psi}^i_{j,l} = (g_i * \psi_{j,l})(x_i)$. We write Eq. (7.10) using the shorter notation

$$\mathbf{\Psi d} = \mathbf{y}. \tag{7.11}$$

with $\mathbf{d}$ covering scaling function and wavelet coefficients.

By definition wavelets fulfill $\int \psi_{j,l} = 0$ and with increasing scale j the $\psi_{j,l}$ become narrower. As a consequence, the convolution with the smoothing kernel $(g * \psi_{j,l})$ will diminish towards zero as j increases. In other words, a sample point's significance on the wavelet coefficients $d_{j,l}$ decreases. At the same time, a coarse scale sample point has less influence on coefficient $d_{j,l}$ than a fine scale sample point at the same position because the convolution kernel g is broader. In this way, we respect all given samples but prevent coarse scale samples from interfering with fine scale surface structures.

7.4 Surface Reconstruction

Samples given in a real application are disturbed by noise. Regions are irregularly sampled regarding not only density but also their scale. The consequence is that the linear system (7.11) cannot be solved exactly and we have to formulate an optimization problem. We introduce and discuss a regularization to avoid over-fitting and show how the problem can be solved efficiently. Thereafter we discuss how the smoothing kernel g_i can be estimated or even influenced in the context of multi-view stereo sample points and examine whether an optimal kernel exists. At the end of this section we review a particular wavelet family which we use in our experiments.

7.4.1 Optimization

The main problem we face when fitting a function to sample points is to reconstruct a smooth surface while still modeling the details. Besides the presence of noise and sparse sampling our model has a more inherent problem of over-fitting. When trying to recover fine scale details that are not sufficiently supported by the data, the entries of an entire row of the matrix $\mathbf{\Psi}$ vanish, and there is almost no control on the corresponding wavelet

coefficients $d_{j,l}$. One way to counteract this is to decrease the maximum scale $j_{\max}$ but this effect might just be local and we do not want to decrease the overall detail level according to the worst represented region. Consequently, a regularization is necessary that prevents all kinds of over-fitting. We add a penalty on the second order derivatives similar to Calakli and Taubin [2011] and solve the following optimization problem

$$\underset{\mathbf{d}}{\text{minimize}} \quad \frac{1}{N}\|\mathbf{\Psi d}-\mathbf{y}\|^2 + \lambda \int \|Hf(x)\|^2 \, dx \tag{7.12}$$

where f denotes the final surface represented as in Eq. 7.4. $Hf(x)$ is the Hessian containing the second-order partial derivatives of f and $\|Hf(x)\|$ is the Frobenius norm of the matrix $Hf(x)$. Note that the smoothing term automatically affects regions with low-scale samples more than regions where high-scale samples are present because the corresponding coefficients are less restricted. We can reformulate the problem into a quadratic program

$$\underset{\mathbf{d}}{\text{minimize}} \quad \mathbf{d}^{\mathrm{T}}[\frac{1}{N}\mathbf{\Psi}^{\mathrm{T}}\mathbf{\Psi} + \lambda Q^s]\mathbf{d} - \frac{2}{N}\mathbf{y}^{\mathrm{T}}\mathbf{\Psi d} \tag{7.13}$$

where the matrix Q^s is the contribution of the second order derivative term. It consists of

$$Q^s_{\alpha,\beta} = \int < H\chi_\alpha(x), H\chi_\beta(x) > \, dx. \tag{7.14}$$

where we used the indices α and β to consecutively number the basis functions χ_α which are either scaling functions or wavelets. The matrix $Q = \frac{1}{N}\mathbf{\Psi}^{\mathrm{T}}\mathbf{\Psi} + \lambda Q^s$ is symmetric and positive definite, so the problem can be solved using a standard quadratic program solver.

7.4.2 Scale Estimation

Until now we assumed that the convolution kernels g_i are known. However, it is not clear how to determine the kernel for given sample points in a real-world application. In the context of patch-based depth reconstruction we provided an approximation of the smoothing kernel in the previous two chapters. We first showed that the window based photo-consistency optimization between images leads to sample points that lie on a box filtered version of the original surface. The width of the box filter can be computed from the pixels footprint, i.e., the projected size of the pixel spacing in world space, multiplied with the window size in pixels.

In the previous chapter, we applied a weighted photo-consistency optimization for depth reconstruction and showed that the convolution kernel is equal to the applied

weighting function (accordingly scaled to match the world-coordinate system). This not only allows us to estimate the convolution kernel g_i for the samples but to actively influence it during creation of the sample points. We will exploit this in our experiments in Sec. 7.5.

7.4.3 Optimal Smoothing Kernel

Before presenting the results of our method we want to spend some extra thought on choosing the optimal smoothing kernel. Ideally, the way the samples are generated matches the multi-resolution analysis used for the surface representation. In other words the significance of a sample point vanishes completely for all wavelet coefficients $d_{j,l}$ with j larger than the sample's scale. How can this be modeled? In the case of (semi-) orthogonal wavelets we have

$$< \phi_{0,k}, \psi_{j,l} > = 0, \text{ for all } j \geq 0. \tag{7.15}$$

If we further assume symmetric scaling functions we can establish the following relationship between the inner product and the convolution

$$< \phi_{0,k}, \psi_{j,l} > = \int \phi(t-k)\psi_{j,l}(t)\,dt \tag{7.16}$$

$$= (\phi * \psi_{j,l})(k) = 0. \tag{7.17}$$

That is, if we had $g_i(t) = \phi(t)$ as the convolution kernel and samples at the integer positions $x_i \in \mathbb{Z}$ we would get

$$y_i = (\phi * f)(x_i) = \sum_l c_{0,l}(\phi * \phi)(x_i + l). \tag{7.18}$$

Having this kind of sample points we could solely solve for the scaling function coefficients $c_{0,l}$. Following this path, with $g_i(t) = \phi(2^j t)$ and sampling positions $x_i \in \{2^{-j}k, k \in \mathbb{Z}\}$ one could obtain the wavelet coefficients up to $d_{j-1,l}$. Note that in such a scenario the inherent over-fitting discussed in Sec. 7.4.1 is removed to a large extent.

Unfortunately, due to obvious reasons this is not achievable in practice: Firstly, we are very likely to not exactly hit the desired sampling positions and secondly we are incapable to (exactly) control the dilation of the smoothing kernel. In addition, we lose the possibility to exploit redundancy by sampling more positions than actually required. Therefore it remains a thought experiment and in practice we prefer to choose a smoothing kernel that behaves well and simplifies computations.

7.4.4 Spline Wavelets on the Interval

We now further specify the surface representation. Because the observed surface will always be of finite extent we can only identify corresponding coefficients. Consequently, there is no point in describing the surface using wavelets on the entire $\mathbb{R}^2$ (or $\mathbb{R}$) which would lead to border handling problems. Therefore we employ wavelets on bounded intervals, w.l.o.g. on $[0,1]$.

For our implementation we decided to use spline wavelets. From a variety of good reasons to do so (see Unser [1997]) we out point two: First, closed form solutions exist, not only for the basis functions but also for the convolution with, e.g., a Gaussian. Second, the basis functions are smooth allowing us to easily represent smooth surfaces. In the following we will shortly review the semi-orthogonal spline wavelets on $L^2([0,1])$ which were initially introduced by Chui and Quak [1992] (see also Stollnitz *et al.* [1996]). They are a natural extension of the semi-orthogonal spline wavelets on $L^2(\mathbb{R})$ developed by Chui and Wang [1992].

A basis for V_j is given by the B-splines $B_{i,m,j}$ with $i = -m+1,\ldots,2^j-1$ which are defined as follows:

$$B_{i,m,j} = (t^{(j)}_{i+m} - t^{(j)}_i)[t^{(j)}_i,\ldots,t^{(j)}_{i+m}]_t (t-x)^{m-1}_+ \tag{7.19}$$

$$t^{(j)}_k = \begin{cases} 0, & k = -m+1,\ldots,0 \\ k2^{-j}, & k = 1,\ldots,2^j-1 \\ 1, & k = 2^j,\ldots,2^j+m-1 \end{cases} \tag{7.20}$$

where m denotes the spline order and the term $[\cdot,\ldots,\cdot]_t$ refers to the m-th divided difference of $(t-x)^{m-1}_+$ with respect to t. The inner scaling functions $B_{i,m,j}$, for $i = 0,\ldots,2^j-m$, are equal to the scaling functions for $L^2(\mathbb{R})$ which are just dilations and translations of the cardinal B-spline $N_m(x) = m[0,1,\ldots,m]_t(t-x)^{m-1}_+$:

$$\phi_{j,i}(x) = B_{i,m,j}(x) = N_m(2^j x - i),\ i = 0,\ldots,2^j-m. \tag{7.21}$$

The inner wavelets are equal to the Chui–Wang wavelets of order m:

$$\psi_{j,i}(x) = \frac{1}{2^{2m-1}} \sum_{k=0}^{2m-2} (-1)^k N_{2m}(k+1) B^{(m)}_{2i+k,2m,t^{(j+1)}_m}(x). \tag{7.22}$$

We refer to Chui and Quak [1992] on how to construct the border wavelets in the general case. For cubic splines ($m = 4$) the coefficients of the refinement equation are given in [Stollnitz et al. 1996, App. B]). Figure 7.3 shows the scaling functions and wavelets for $j = 2$.

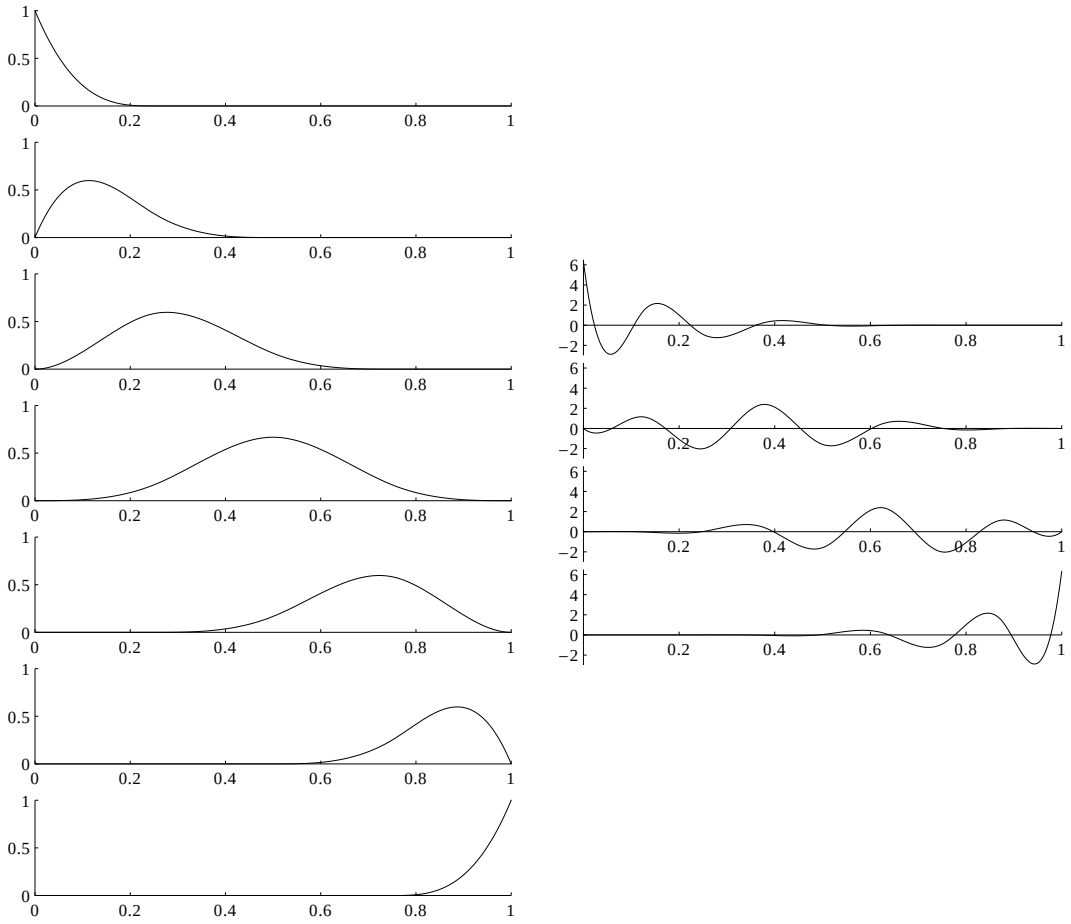


Figure 7.3: The seven scaling functions (*left*) and four wavelets (*right*) on the interval spanning $\mathbf{V}_3$ ($j = 2$).

7.5 Results

For the implementation we use the large-scale optimization software Mosek [Mosek ApS 2012] to solve the optimization problem. In all experiments we assume that the final surface can be described as a height field $z = f(x, y)$ with $(x, y) \in [0, 1]^2$. This is realized using a rigid transformation plus an additional scaling, thus easily invertible after reconstruction.

7.5.1 Synthetic Data

We start with a synthetic data set where we know both the ground truth surface (see Fig. 7.4 (*left*)) and its wavelet decomposition. The input to our method are sample points from the convolved version of this surface using a Gaussian with known standard

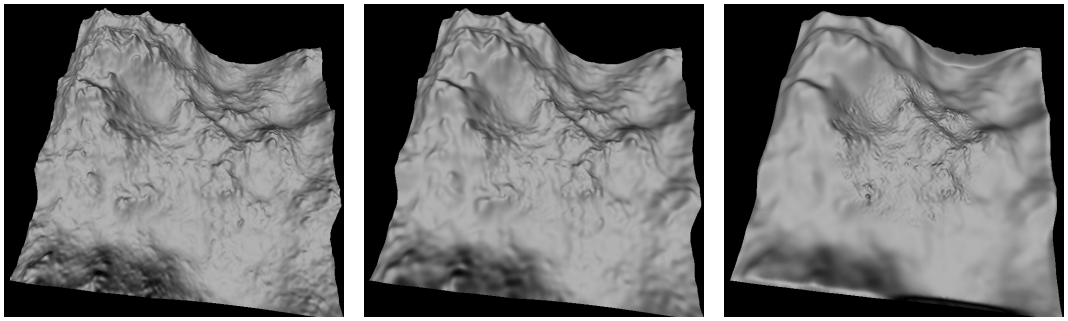


Figure 7.4: *Left:* Ground truth surface from which we generate low- and high-scale samples. *Middle:* Our reconstruction taking scale into account. *Right:* Treating all samples as real point samples neglecting the scale.

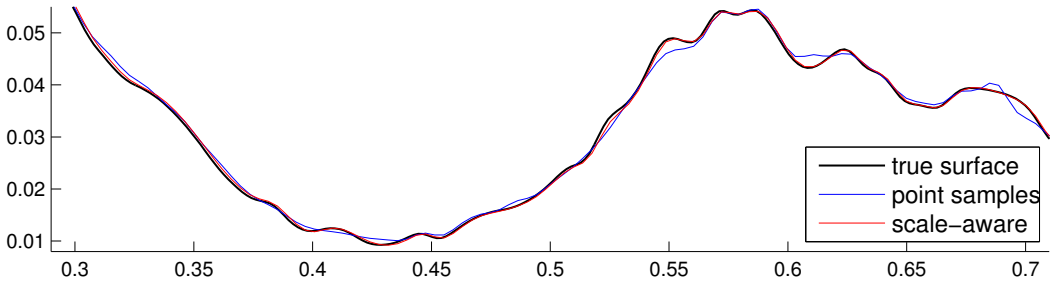


Figure 7.5: A segment of the central horizontal scanline through the geometry in Fig. 7.4 showing that our scale-aware reconstruction accurately follows the ground truth.

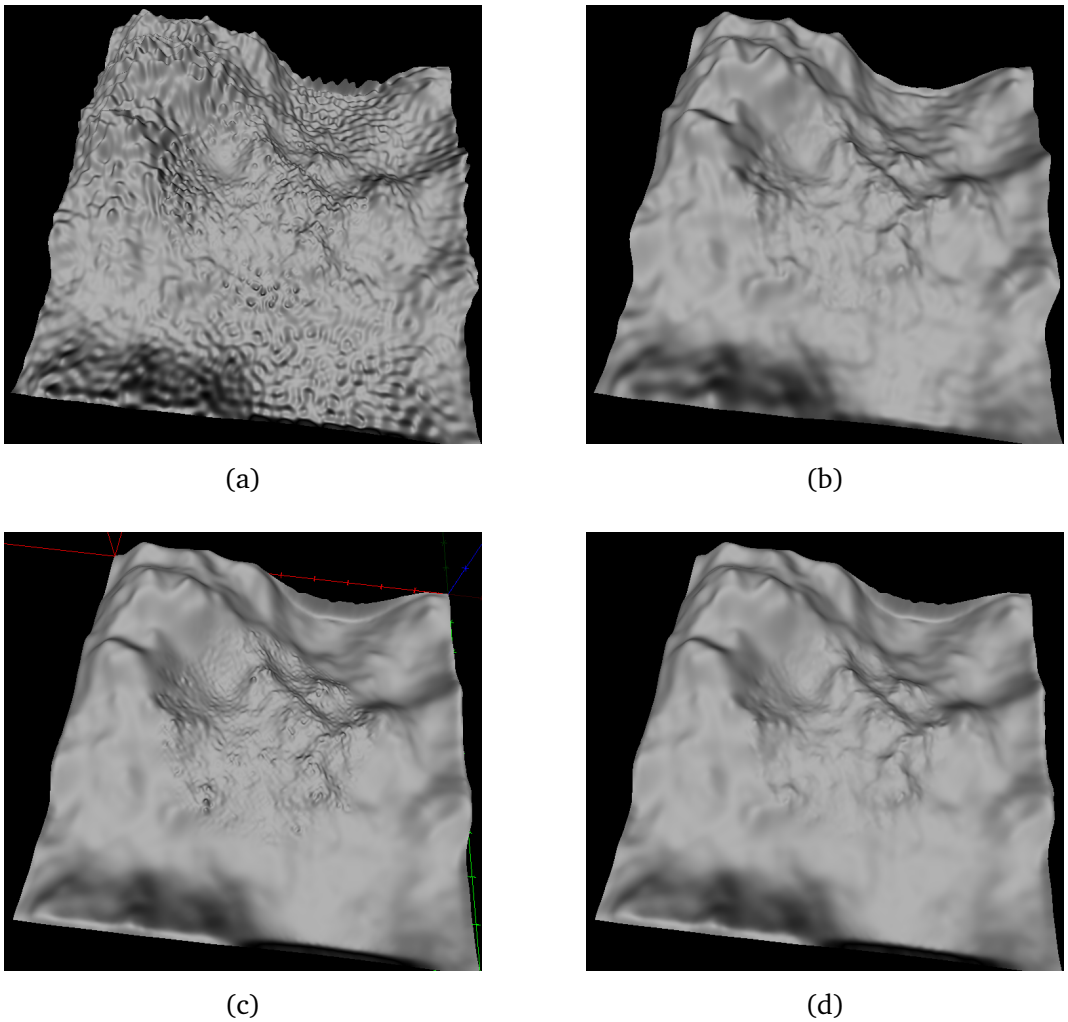
deviation σ. We generate 20,000 sample points from which $\frac{4}{5}$ are uniformly sampled over $[0,1]^2$ with $\sigma = 0.01$ (low-scale), and $\frac{1}{5}$ are uniformly sampled on a centered circle with radius 0.25 with $\sigma = 0.002$ (high-scale). For the reconstruction we use $j_0 = 4$ and $j_{\max} = 6$, i.e., using scaling functions $\phi_{4,\cdot}$ and wavelets $\psi_{4,\cdot}, \psi_{5,\cdot}, \psi_{6,\cdot}$. The smoothness weight is $\lambda = 10^{-12}$. The result of our method can be seen in Fig. 7.4 (*middle*). In Fig. 7.4 (*right*) we assumed all input samples are real point samples which means that g is the Dirac delta function. The benefit of taking the scale into account, even in the areas with only low-scale sample points, is clearly visible. Fig. 7.5 shows a segment from the center horizontal scanline that confirms this impression.

In Fig. 7.6 we demonstrate the effect of the smoothness weight. We reconstruct effectively on the same scale, that is in V_7, but using scaling functions $\phi_{6,\cdot}$ and wavelets $\psi_{6,\cdot}$. Now, the smoothing kernel is roughly as big as the basis function and there is only very small or no data force on the basis function coefficients leading to "ripple" artifacts. The same effect can be caused by under-sampling. Then the smoothness weight λ has to be chosen accordingly to prevent introducing high-frequency artifacts.

7.5.2 Real-World Data

To test our algorithm on real-world data we took 174 images of a relief on a stone wall (see Fig. 7.7). We registered the images using structure-from-motion [Snavely et al. 2008] and reconstructed depth maps per view using a multi-view stereo implementation similar to Goesele *et al.* [2007]. In contrast to them we use a weighted photo-consistency optimization. More precisely we use a patch of size 21×21 pixels in image space and apply a Gaussian with $\sigma = 4$. We use such a big patch to get less noise in the reconstruction and to achieve a reasonably sized smoothing kernel to better visualize the effect of our method. The input images have a resolution of about 1000×666 pixels. According to our results in the previous chapter we can then estimate the smoothing kernel g to be a Gaussian as well, with a scaled standard deviation depending on the internal camera parameters and the estimated depth. In order to meet the height field assumption we fit a plane to the feature points obtained by structure-from-motion and compute a transformation that maps it on the x, y—plane. As input to our method we merge the reconstructed points from 6 depth maps covering a range of about factor 3 in scale, i.e., $\sigma_{\max} \approx 3\sigma_{\min}$. This yields a total of about 1.6 million points.

We reconstruct a surface using $j_0 = 5$ and $j_{\max} = 6$, i.e., using $35^2 = 1,225$ scaling functions $\phi_{5,\cdot}$ spanning V_5, 3,264 wavelets $\psi_{5,\cdot}$ spanning W_5, and 12,672 wavelets $\psi_{6,\cdot}$ spanning W_6. In total we optimize for 17,161 basis function coefficients. Fig. 7.8 shows the comparison between our scale-aware (*left*) reconstruction and using the same setup

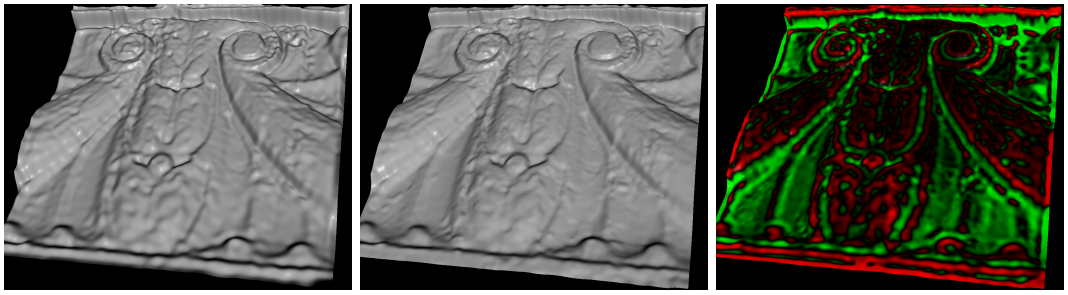


(a) (b)

(c) (d)

Figure 7.6: The starting scale $j_0 = 6$ is chosen higher than in Fig. 7.4 resulting in less supported coefficients of the scaling function. *(a)* A small smoothness weight ($\lambda = 10^{-12}$) can lead to artifacts. *(b)* Choosing a larger weight ($\lambda = 10^{-10}$) fixes this problem. *(c)+(d)* Using the same smoothness weights ($\lambda = 10^{-12}$ and $\lambda = 10^{-10}$, respectively) but assuming all samples are real point samples. This variant is naturally less sensitive to the smoothness weight but also preserves less detail.

Figure 7.7: Example input images of the Relief data set.



(a) (b) (c)

Figure 7.8: Reconstruction using $j_0 = 5$ and $j_{max} = 6$. *(a)* Taking scale into account preserves more detail compared to treating all samples as real point samples in *(b)*. The colored mesh *(c)* has vertex positions identical to *(b)* and the vertex colors encode the differences in height compared to *(a)*. Changes mainly affect the edges since we amplify high frequencies.

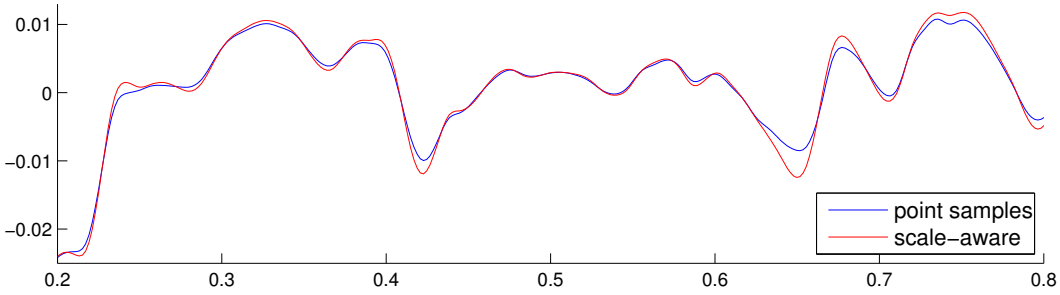


Figure 7.9: A profile of the Relief reconstruction (see Fig. 7.8) showing that our scale-aware reconstruction preserves more detail than treating all samples as real point samples.

but ignoring scale (*center*), i.e., treating all samples as real point samples. Detail in the middle and lower part of the rendering is emphasized while some artifacts from multi-view stereo become more visible.

7.6 Discussion

We present a general surface reconstruction framework that incorporates the (multi-) scale property of the samples points. To our knowledge we are the first to dissolve the paradigm of point samples that lie on the true surface but still incorporate all data in the reconstruction process. Using the concept of multi-resolution analysis we can merge the sample points in frequency space while still maintaining locality due to the wavelet

basis. On synthetic data we demonstrate clearly that our method correctly integrates the multi-scale input data. The real-world example indicates the improvement of our method as well, however, we have to struggle with registration errors and multi-view stereo artifacts. As already pointed out in previous chapters the modeling of the multi-view stereo reconstruction is imperfect and thus the estimated smoothing kernel is not accurate. Experience from the image domain (e.g. [Levin et al. 2011]) suggests that a better kernel estimate will likely improve reconstruction quality.

The biggest limitation of our method is probably the current restriction to height fields. Using an implicit surface representation, e.g., the signed distance field, would allow to extend the method to a more general class of surfaces. We do, however, face the problem that it is still unclear how reconstruction techniques affect the signed distance field.

8 Conclusions and Future Research Directions

Contents

SURFACE reconstruction from sample points has to face new challenges. The application has shifted from the reconstruction of single objects to entire scenes, regions are represented with varying sampling density and sample points even represent different scales. Naively fusing these points can suppress details or introduce high-frequency artifacts at regions where fine- and coarse-scale samples meet. This thesis presented two new surface reconstruction algorithms that handle multi-scale input data in distinct ways. We also answer the question to what extent sample points originating from patch-based matching between images can preserve details or smoothe the surface, respectively. Understanding the characteristics of the multi-scale input points is a vital ingredient in order to develop accurate surface reconstruction algorithms. The next section summarizes the main contributions of the thesis and is followed by a high-level discussion with an outlook to possible future research directions.

8.1 Summarizing Contributions

The work presented in this thesis contributes to both the field of multi-view stereo and surface reconstruction. At this point we review the contributions listed in Section 1.3 of the introduction with a focus on usability and benefit for the research community.

In Chapter 4 we presented a robust surface reconstruction method that proved to be applicable on a wide range of data sets. It achieves top results on a benchmark data set but includes the notion of the *footprint* of a sample to improve the performance on multi-scale data sets. We model the intuition that fine-scale samples better capture surface details than coarse-scale samples and spread the uncertainty in space accordingly. Together with the hierarchical structure we are able to process large-scale data sets with drastically different sampling rates and level of detail. The resulting triangle meshes are manifold and watertight. Additionally, they feature an adaptive triangulation with smaller triangles in high-detailed regions. The proposed method supersedes the traditionally used methods such as Poisson surface reconstruction at the end of our reconstruction pipeline (see Figure 3.1). The source code is available online [Mücke et al. 2012] and can easily be used by others as well.

In order to better understand the multi-scale characteristics of our input points we proposed a geometrical model that describes the patch-based multi-view stereo reconstruction process (Chapter 5). This model allows us to theoretically analyze the reconstruction accuracy using standard mathematical tools from signal processing. We prove that our model fulfills the linear system properties and determine the modulation transfer function which describes how details are recovered in relation to the patch size. Experiments on synthetic and real-world data sets using a popular multi-view stereo algorithm validate the credibility of our theoretic geometrical model. Our results clearly show a significant amplitude loss of high frequencies in accordance to our model within the limitations of registration errors and inaccuracies of the photo-consistency optimization. This loss of detail has not been modeled or described before and consequently not been considered by any surface reconstruction algorithm. Based on our results we question the common assumption that samples from multi-view stereo are true surface samples that are just disturbed by zero-mean noise and think that our insights have the potential to steer future research in the area of multi-scale surface reconstruction.

Our analysis of patch-based multi-view stereo revealed systematic high-frequency artifacts, basically caused by amplitude inversion. Motivated by this bad frequency behavior we proposed a weighted patch fitting which maps to a generalized reconstruction model. We proved that under common criteria for the weighting function the reconstructed surface is a convolution of the original surface with a dilated version of the weighting function. At the same time, we hereby shifted the earlier formulation in frequency space to a locally evaluable convolution in geometry space. The derived criteria for the weighting function allow for choosing a wide range of filters for patch-based reconstruction and thus for various convolution kernels. In particular, true low-pass filters,

e.g. a Gaussian, can be realized removing the high-frequency artifacts. As we showed for a particular multi-view stereo algorithm it is easy to implement the weighted patch fitting, so already existing methods (*e.g.* [Furukawa and Ponce 2010]) can benefit from our insights. In analogy to super-resolution methods in the image domain one can think of our results as providing the generative model of patch-based depth reconstruction. This could be used to apply the various methods and the broad knowledge from the image domain to depth map recovery, *e.g.*, from multiple weighted reconstructions.

Finally, we presented a general surface reconstruction framework for 2.5D surfaces that incorporates our insights about multi-scale sample points. To our knowledge, this is the first approach that abandons the paradigm that the input points are samples from the true surface disturbed by zero-mean noise while still taking all available data into account. By representing the final surface in a wavelet domain we obtain a space-frequency decomposition and can compute the influence of a sample point to each space-frequency window. Our framework works for various kinds of multi-scale input data and allows to characterize each sample point with a different convolution kernel. The entire surface reconstruction problem boils down to a quadratic program that can be minimized by solving the corresponding linear system. The benefit of our method is clearly visible on synthetic data where it is clearly indispensable to take the scale information into account. Regarding real-world data the positive effects are attenuated by registration errors and artifacts introduced by the photo-consistency optimization. To overcome the limitation to 2.5D surfaces an implicit surface representation such as a signed distance field could be used along with our framework. The missing ingredient is to determine a model for the smoothed variants of the surface that maps the multi-scale sample points.

8.2 Discussion and Future Work

Present multi-view stereo algorithms already achieve very accurate results [Seitz et al. 2006] and have been successfully applied to reconstruct real-world objects or even large scenes [Furukawa et al. 2010]. There are also methods that generate a closed surface in the form of a triangle mesh, *e.g.* the depth map fusion by Fuhrmann and Goesele [2011] or the algorithm presented in Chapter 4. Geometry reconstruction from images is consequently already a good alternative to active capture devices and works in practice. In the following, we suggest future research directions inspired by the content of this thesis and point out the main paradigm shift we infer from the presented research results.

8.2.1 Analysis of Multi-View Stereo

In order to develop tailored surface reconstruction algorithms we are convinced that a better understanding of the reconstruction (or registration) errors is necessary. This includes studying the impact of the surface texture, parallax between the views, and angle between the epipolar lines on the reconstruction accuracy. It is also worth to investigate how far photo-consistency optimization is correctly modeled by the least squares planar patch fit as we proposed in this thesis. Does this model still hold for object boundaries or depth discontinuities? How robust is it against different photo-consistency measures [Hu and Mordohai 2012]? As our surface reconstruction results indicated it is highly beneficial to correctly model the sample points' systematic error and uncertainty distribution. Additional insight might also influence and improve multi-view stereo reconstruction techniques. We have seen such an example regarding the proposed weighted patch fit demanding for a weighted photo-consistency optimization.

8.2.2 Surface Reconstruction

Despite a huge amount of existing work feature-preserving surface reconstruction is an ongoing research topic [Berger et al. 2013, Kazhdan and Hoppe 2013]. In this thesis we showed that it is important to take the scale characteristics of the sample points into account. This can be interpreted as modeling a systematic error present in the input data. Along these lines it is worth to model and investigate other potential multi-view stereo reconstruction errors or noise characteristics. In both presented surface reconstruction algorithms, as in many other methods, it is easily possible to consider confidence values. In the case of multi-view stereo, however, meaningful confidence values are hard to assign. Especially in the surface reconstruction from depth maps there is a lot of redundant data where error characteristics and meaningful confidence values can be exploited to correctly recover the original surface. Besides better modeling the error characteristics of the sample points a sensible regularization term has proven to be extremely beneficial. In fact, without a regularization term the surface reconstruction problem is effectively intractable. In analogy to the image domain, where natural image priors have been developed and successfully applied, surface reconstruction might benefit from more sophisticated or even application specific priors.

8.2.3 Paradigm Shift

Geometry reconstruction from images can clearly benefit from advancement in the areas of multi-view stereo or in the area of surface reconstruction from sample points.

However, the results presented in this thesis revealed that additionally a widely and maybe unconsciously used paradigm has to be reconsidered. This paradigm can be essentially phrased as follows: Increasing the number of images and thus the number of reconstructed multi-view stereo points (*e.g.*, by reconstructing depth maps) will eventually result in the perfect surface. The underlying assumption is that the sample points obtained by multi-view stereo are point samples from the true surface disturbed by zero-mean noise and some outliers. Surface reconstruction algorithms (*e.g.*, VRIP [Curless and Levoy 1996]) can remove these errors and the reconstruction quality improves with an increasing number of sample points. The results of Chapter 5 and 6 show that there is an additional systematic error mainly effecting the fine details. The consequence is that with an increasing number of images and depth maps the reconstruction quality does not improve automatically, instead it can even deteriorate if this systematic error is not considered.

The thesis also introduced means to model the systematic error using a modulation transfer function (Chapter 5) and in terms of a convolution (Chapter 6), respectively. It is still not obvious how to correct for this error. Since the modulation transfer function has multiple zeros it is not invertible and the deconvolution is not uniquely solvable for general filters. In practice, however, we can try to compute meaningful approximations. For example, the geometry can be recovered up to a certain frequency supported by the reconstructed sample points or one can apply surface priors. The topic of deconvolution is well researched in signal and image processing and can probably inspire future surface reconstruction algorithms. In our opinion, the crucial point is how to incorporate the knowledge gained about depth maps in order to obtain the perfect 3D reconstruction. The algorithm presented in Chapter 7, despite its current limitation to 2.5D surfaces, constitutes a starting point and abandons the aforementioned paradigm.

Bibliography

AGARWAL, S., SNAVELY, N., SIMON, I., SEITZ, S. M., AND SZELISKI, R. 2009. Building Rome in a day. In *Proc. of International Conference on Computer Vision (ICCV)*, IEEE Computer Society.

ALEXA, M., BEHR, J., COHEN-OR, D., FLEISHMAN, S., LEVIN, D., AND SILVA, C. T. 2003. Computing and rendering point set surfaces. *Visualization and Computer Graphics 9*, 1, 3–15.

ALLIEZ, P., COHEN-STEINER, D., TONG, Y., AND DESBRUN, M. 2007. Voronoi-based variational reconstruction of unoriented point sets. In *Proc. of Eurographics Symposium on Geometry Processing (SGP)*, Eurographics, 39–48.

AMENTA, N., AND BERN, M. 1999. Surface reconstruction by voronoi filtering. *Discrete & Computational Geometry 22*, 4, 481–504.

AMENTA, N., CHOI, S., AND KOLLURI, R. K. 2001. The power crust, unions of balls, and the medial axis transform. *Computational Geometry 19*, 2–3, 127–153.

AMENTA, N., CHOI, S., DEY, T. K., AND LEEKHA, N. 2002. A simple algorithm for homeomorphic surface reconstruction. *International Journal of Computational Geometry & Applications 12*, 1-2, 125–141.

BAILER, C., FINCKH, M., AND LENSCH, H. P. 2012. Scale robust multi view stereo. In *Proc. of European Conference on Computer Vision (ECCV)*, vol. 7574 of *Lecture Notes in Computer Science (LNCS)*. Springer Berlin Heidelberg, 398–411.

BELLOCCHIO, F., BORGHESE, N. A., FERRARI, S., AND PIURI, V. 2013. *3D Surface Reconstruction: Multi-Scale Hierarchical Approaches*. Springer, New York, NY.

BERGER, M., LEVINE, J. A., NONATO, L. G., TAUBIN, G., AND SILVA, C. T. 2013. A benchmark for surface reconstruction. *Transactions on Graphics (TOG) 32*, 2, 20:1–20:17.

BOISSONNAT, J.-D. 1984. Geometric structures for three-dimensional shape representation. *Transactions on Graphics (TOG) 3*, 4, 266–286.

BOLITHO, M., KAZHDAN, M., BURNS, R., AND HOPPE, H. 2007. Multilevel streaming for out-of-core surface reconstruction. In *Proc. of Eurographics Symposium on Geometry Processing (SGP)*, Eurographics, 69–78.

BOYKOV, Y., AND KOLMOGOROV, V. 2003. Computing geodesics and minimal surfaces via graph cuts. In *Proc. of International Conference on Computer Vision (ICCV)*, IEEE Computer Society, 26–33.

BOYKOV, Y., AND KOLMOGOROV, V. 2004. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *Transactions on Pattern Analysis and Machine Intelligence (PAMI) 26*, 9, 1124–1137.

BOYKOV, Y., AND VEKSLER, O. 2006. Graph cuts in vision and graphics: Theories and applications. In *Handbook of Mathematical Models in Computer Vision*. Springer.

BOYKOV, Y., VEKSLER, O., AND ZABIH, R. 2001. Fast approximate energy minimization via graph cuts. *Transactions on Pattern Analysis and Machine Intelligence (PAMI) 23*, 11, 1222–1239.

BRADLEY, D., BOUBEKEUR, T., AND HEIDRICH, W. 2008. Accurate multi-view reconstruction using robust binocular stereo and surface meshing. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society.

CALAKLI, F., AND TAUBIN, G. 2011. SSD: Smooth signed distance surface reconstruction. *Computer Graphics Forum (CGF) 30*, 7, 1993–2002.

CARR, J. C., BEATSON, R. K., CHERRIE, J. B., MITCHELL, T. J., FRIGHT, W. R., MCCALLUM, B. C., AND EVANS, T. R. 2001. Reconstruction and representation of 3D objects with radial basis functions. In *Proc. of SIGGRAPH*, ACM, New York, NY, USA, 67–76.

CAZALS, F., AND GIESEN, J. 2006. Delaunay triangulation based surface reconstruction. In *Effective Computational Geometry for Curves and Surfaces*, J.-D. Boissonnat and M. Teillaud, Eds. Springer Berlin Heidelberg, 231–276.

CHUI, C. K., AND QUAK, E. 1992. Wavelets on a bounded interval. In *Numerical Methods in Approximation Theory*, D. Braess and L. L. Schumaker, Eds., vol. 105 of *ISNM 105: International Series of Numerical Mathematics*. Birkhäuser Basel, 53–75.

CHUI, C. K., AND WANG, J. Z. 1992. On compactly supported spline wavelets and a duality principle. In *Transactions of the American Mathematical Society*.

COHEN, L. D., AND COHEN, I. 1993. Finite-element methods for active contour models and balloons for 2-D and 3-D images. *Transactions on Pattern Analysis and Machine Intelligence (PAMI) 15*, 11, 1131–1147.

COHEN-STEINER, D., AND DA, F. 2004. A greedy Delaunay-based surface reconstruction algorithm. *The Visual Computer 20*, 1, 4–16.

CURLESS, B., AND LEVOY, M. 1996. A volumetric method for building complex models from range images. In *Proc. of SIGGRAPH*, ACM, 303–312.

DEY, T. K., AND GOSWAMI, S. 2003. Tight cocone: a water-tight surface reconstructor. In *Proc. of Symposium on Solid Modeling and Applications*, ACM, New York, NY, USA, SM '03, 127–134.

DEY, T. K., LI, K., RAMOS, E. A., AND WENGER, R. 2009. Isotopic reconstruction of surfaces with boundaries. *Computer Graphics Forum (CGF) 28*, 5, 1371–1382.

DIGNE, J., MOREL, J. M., SOUZANI, C.-M., AND LARTIGUE, C. 2011. Scale space meshing of raw data point sets. *Computer Graphics Forum (CGF) 30*, 6, 1630–1642.

DOI, A., AND KOIDE, A. 1991. An efficient method of triangulating equi-valued surfaces by using tetrahedral cells. *IEICE Transactions on Information and Systems E74-D*, 1, 214–224.

DONG, B., AND SHEN, Z. 2011. Wavelet frame based surface reconstruction from unorganized points. *Computational Physics 230*, 22, 8247–8255.

ESTEVE, J., BRUNET, P., AND ÀLVAR VINACUA. 2005. Approximation of a variable density cloud of points by shrinking a discrete membrane. *Computer Graphics Forum (CGF) 24*, 4, 791–807.

FLEISHMAN, S., COHEN-OR, D., AND SILVA, C. T. 2005. Robust moving least-squares fitting with sharp features. In *Proc. of SIGGRAPH*, ACM, New York, NY, USA, 544–552.

FRAHM, J.-M., FITE-GEORGEL, P., GALLUP, D., JOHNSON, T., RAGURAM, R., WU, C., JEN, Y.-H., DUNN, E., LAZEBNIK, S., AND POLLEFEYS, M. 2010. Building Rome on a cloudless day. In *Proc. of European Conference on Computer Vision (ECCV)*, Springer Berlin Heidelberg, vol. 6314 of *Lecture Notes in Computer Science (LNCS)*, 368–381.

FUHRMANN, S., AND GOESELE, M. 2011. Fusion of depth maps with multiple scales. In *Proc. of SIGGRAPH Asia*, ACM, New York, NY, USA, 148:1–148:8.

FURUKAWA, Y., AND PONCE, J. 2010. Accurate, dense, and robust multi-view stereopsis. *Transactions on Pattern Analysis and Machine Intelligence (PAMI) 32*, 8, 1362–1376.

FURUKAWA, Y., CURLESS, B., SEITZ, S. M., AND SZELISKI, R. 2009. Manhattan-world stereo. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society.

FURUKAWA, Y., CURLESS, B., SEITZ, S. M., AND SZELISKI, R. 2010. Towards Internet-scale multi-view stereo. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, 1434–1441.

GAL, R., SHAMIR, A., HASSNER, T., PAULY, M., AND COHEN-OR, D. 2007. Surface reconstruction using local shape priors. In *Proc. of Eurographics Symposium on Geometry Processing (SGP)*, Eurographics, 253–262.

GALLUP, D., FRAHM, J.-M., AND POLLEFEYS, M. 2010. Piecewise planar and non-planar stereo for urban scene reconstruction. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, 1418–1425.

GARGALLO, P., AND STURM, P. 2005. Bayesian 3D modeling from images using multiple depth maps. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society.

GARGALLO, P., PRADOS, E., AND STURM, P. 2007. Minimizing the reprojection error in surface reconstruction from images. In *Proc. of International Conference on Computer Vision (ICCV)*, IEEE Computer Society, 1–8.

GOESELE, M., FUCHS, C., AND SEIDEL, H.-P. 2003. Accuracy of 3D range scanners by measurement of the slanted edge modulation transfer function. In *Proc. of International Conference on 3-D Digital Imaging and Modeling (3DIM)*, IEEE Computer Society.

GOESELE, M., SNAVELY, N., CURLESS, B., HOPPE, H., AND SEITZ, S. M. 2007. Multi-view stereo for community photo collections. In *Proc. of International Conference on Computer Vision (ICCV)*, IEEE Computer Society.

GOLDLUECKE, B., AND CREMERS, D. 2009. A superresolution framework for high-accuracy multiview reconstruction. In *DAGM*, vol. 5748 of *Lecture Notes in Computer Science (LNCS)*.

GRUEN, A., AND BALTSAVIAS, E. P. 1988. Geometrically constrained multiphoto matching. *Photogrammetric Engineering & Remote Sensing 54*, 5, 633–641.

HABBECKE, M., AND KOBBELT, L. 2007. A surface-growing approach to multi-view stereo reconstruction. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society.

HARRIS, C., AND STEPHENS, M. 1988. A combined corner and edge detector. In *Proc. of the Alvey Vision Conference*, Alvety Vision Club, 23.1–23.6.

HERNÁNDEZ ESTEBAN, C., AND SCHMITT, F. 2004. Silhouette and stereo fusion for 3D object modeling. *Computer Vision and Image Understanding 96*, 3, 367–392.

HIEP, V., KERIVEN, R., LABATUT, P., AND PONS, J.-P. 2009. Towards high-resolution large-scale multi-view stereo. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society.

HOPPE, H., DEROSE, T., DUCHAMP, T., MCDONALD, J., AND STUETZLE, W. 1992. Surface reconstruction from unorganized points. In *Proc. of SIGGRAPH*, vol. 26, ACM, 71–78.

HORNUNG, A., AND KOBBELT, L. 2006. Hierarchical volumetric multi-view stereo reconstruction of manifold surfaces based on dual graph embedding. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society.

HORNUNG, A., AND KOBBELT, L. 2006. Robust reconstruction of watertight 3D models from non-uniformly sampled point clouds without normal information. In *Proc. of Eurographics Symposium on Geometry Processing (SGP)*, Eurographics.

HOSNI, A., BLEYER, M., GELAUTZ, M., AND RHEMANN, C. 2009. Local stereo matching using geodesic support weights. In *Proc. of International Conference on Image Processing (ICIP)*.

HU, X., AND MORDOHAI, P. 2012. A quantitative evaluation of confidence measures for stereo vision. *Transactions on Pattern Analysis and Machine Intelligence (PAMI) 34*, 11, 2121–2133.

JANCOSEK, M., SHEKHOVTSOV, A., AND PAJDLA, T. 2009. Scalable multi-view stereo. In *Proc. of International Conference on Computer Vision Workshops (ICCV Workshops)*, IEEE Computer Society, 1526–1533.

JENKE, P., WAND, M., BOKELOH, M., SCHILLING, A., AND STRASSER, W. 2006. Bayesian point cloud reconstruction. *Computer Graphics Forum (CGF) 25*, 3, 379–388.

JI, H., SHEN, Z., AND XU, Y. 2010. Wavelet frame based scene reconstruction from range data. *Journal of Computational Physics 229*, 6, 2093–2108.

JOHNSON, M. J., SHEN, Z., AND XU, Y. 2009. Scattered data reconstruction by regularization in b-spline and associated wavelet spaces. In *Approximation Theory*.

KANADE, T., AND OKUTOMI, M. 1994. A stereo matching algorithm with an adaptive window: Theory and experiment. *Transactions on Pattern Analysis and Machine Intelligence (PAMI) 16*, 9, 920–932.

KAZHDAN, M., AND HOPPE, H. 2013. Screened poisson surface reconstruction. In *Proc. of SIGGRAPH*, vol. 32, ACM, 29:1–29:13.

KAZHDAN, M., BOLITHO, M., AND HOPPE, H. 2006. Poisson surface reconstruction. In *Proc. of Eurographics Symposium on Geometry Processing (SGP)*, Eurographics.

KAZHDAN, M. 2005. Reconstruction of solid models from oriented point sets. In *Proc. of Eurographics Symposium on Geometry Processing (SGP)*, Eurographics.

KLOWSKY, R., AND GOESELE, M. 2013. Wavelet-based surface reconstruction from multiscale sample points. Tech. Rep. 13rp006-GRIS, Dept. of Computer Science, Technische Universität Darmstadt.

KLOWSKY, R., KUIJPER, A., AND GOESELE, M. 2012. Modulation transfer function of patch-based stereo systems. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, 1386–1393.

KLOWSKY, R., MÜCKE, P., AND GOESELE, M. 2012. Hierarchical surface reconstruction from multi-resolution point samples. In *Outdoor and Large-Scale Real-World Scene Analysis*, F. Dellaert, J.-M. Frahm, M. Pollefeys, L. Leal-Taixé, and B. Rosenhahn, Eds., vol. 7474 of *Lecture Notes in Computer Science (LNCS)*. Springer Berlin Heidelberg, 398–418.

KLOWSKY, R., KUIJPER, A., AND GOESELE, M. 2013. Weighted patch-based reconstruction: Linking (multi-view) stereo to scale space. In *SSVM*, A. Kuijper, K. Bredies, T. Pock, and H. Bischof, Eds., vol. 7893 of *Lecture Notes in Computer Science (LNCS)*. Springer Berlin Heidelberg, 234–245.

KOBBELT, L., CAMPAGNA, S., VORSATZ, J., AND SEIDEL, H.-P. 1998. Interactive multiresolution modeling on arbitrary meshes. In *Proc. of SIGGRAPH*, ACM, 105–114.

KOLMOGOROV, V., AND ZABIH, R. 2002. Multi-camera scene reconstruction via graph cuts. In *Proc. of European Conference on Computer Vision (ECCV)*, Springer Berlin Heidelberg, vol. 2352 of *Lecture Notes in Computer Science (LNCS)*, 82–96.

KUTULAKOS, K. N. 2000. Approximate n-view stereo. In *Proc. of European Conference on Computer Vision (ECCV)*, Springer Berlin Heidelberg, vol. 1842 of *Lecture Notes in Computer Science (LNCS)*, 67–83.

LABATUT, P., PONS, J.-P., AND KERIVEN, R. 2007. Efficient multi-view reconstruction of large-scale scenes using interest points, Delaunay triangulation and graph cuts. In *Proc. of International Conference on Computer Vision (ICCV)*, IEEE Computer Society.

LABATUT, P., PONS, J.-P., AND KERIVEN, R. 2009. Robust and efficient surface reconstruction from range data. *Computer Graphics Forum (CGF) 28*, 8, 2275–2290.

LEMPITSKY, V., AND BOYKOV, Y. 2007. Global optimization for shape fitting. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society.

LEVIN, A., WEISS, Y., DURAND, F., AND FREEMAN, W. 2011. Understanding blind deconvolution algorithms. *Transactions on Pattern Analysis and Machine Intelligence (PAMI) 33*, 12, 2354–2367.

LEVIN, D. 2004. Mesh-independent surface interpolation. In *Geometric Modeling for Scientific Visualization*, G. Brunnett, B. Hamann, H. Müller, and L. Linsen, Eds., Mathematics and Visualization. Springer Berlin Heidelberg, 37–49.

LINDEBERG, T. 1994. *Scale-Space Theory In Computer Vision*. Kluwer Academic Publishers.

LORENSEN, W. E., AND CLINE, H. E. 1987. Marching cubes: A high resolution 3D surface construction algorithm. In *Proc. of SIGGRAPH*, ACM, 163–169.

LOWE, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV) 60*, 2, 91–110.

MALLAT, S. 2008. *A Wavelet Tour of Signal Processing*, 3rd ed. Academic Press.

MANSON, J., AND SCHAEFER, S. 2010. Isosurfaces over simplicial partitions of multiresolution grids. In *Proc. of Eurographics*, Eurographics.

MANSON, J., PETROVA, G., AND SCHAEFER, S. 2008. Streaming surface reconstruction using wavelets. In *Proc. of Eurographics Symposium on Geometry Processing (SGP)*.

MICUSIK, B., AND KOSECKA, J. 2008. Multi-view superpixel stereo in man-made environments. Tech. rep., Dept. Computer Science, George Mason University.

MOSEK APS, 2012. mosek. http://mosek.com/products/mosek/.

MÜCKE, P., KLOWSKY, R., AND GOESELE, M. 2011. Surface reconstruction from multi-resolution sample points. In *Proc. of Vision, Modeling, and Visualization Workshop (VMV)*, Eurographics, 105–112.

MÜCKE, P., KLOWSKY, R., AND GOESELE, M., 2012. Surface reconstruction from multi-resolution sample points. http://www.gris.tu-darmstadt.de/projects/multires-surface-recon/.

NAGAI, Y., OHTAKE, Y., AND SUZUKI, H. 2009. Smoothing of partition of unity implicit surfaces for noise robust surface reconstruction. *Computer Graphics Forum (CGF) 28*, 5, 1339–1348.

OHTAKE, Y., BELYAEV, A., ALEXA, M., TURK, G., AND SEIDEL, H.-P. 2003. Multi-level partition of unity implicits. In *Proc. of SIGGRAPH*, 463–470.

OHTAKE, Y., BELYAEV, A., AND SEIDEL, H.-P. 2003. A multi-scale approach to 3D scattered data interpolation with compactly supported basis functions. In *Proc. of Shape Modeling International (SMI)*, IEEE Computer Society, 153–161.

PASTOR, L., AND RODRÍGUEZ, A. 1999. Surface approximation of 3D objects from irregularly sampled clouds of 3d points using spherical wavelets. In *Proc. of International Conference on Image Analysis and Processing (ICIAP)*, IEEE Computer Society.

PAULY, M., KOBBELT, L. P., AND GROSS, M. 2006. Point-based multiscale surface representation. In *Transactions on Graphics (TOG)*, vol. 25, ACM, 177–193.

PHARR, M., AND HUMPHREYS, G., 2012. Physically based rendering. http://pbrt.org.

PICKUP, L., CAPEL, D., ROBERTS, S., AND ZISSERMAN, A. 2007. Bayesian methods for image super-resolution. *The Computer Journal 52*, 1, 101–113.

PONS, J.-P., KERIVEN, R., AND FAUGERAS, O. 2007. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. *International Journal of Computer Vision (IJCV) 72*, 2, 179–193.

REICHENBACH, S. E., PARK, S. K., AND NARAYANSWAMY, R. 1991. Characterizing digital image acquisition devices. *Optical Engineering 30*, 2, 170–177.

SCHAEFER, S., JU, T., AND WARREN, J. 2007. Manifold dual contouring. *Transactions on Visualization and Computer Graphics 13*, 3, 610–619.

SEITZ, S. M., AND DYER, C. R. 1999. Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision (IJCV) 35*, 2, 151–173.

SEITZ, S. M., CURLESS, B., DIEBEL, J., SCHARSTEIN, D., AND SZELISKI, R. 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society.

SEITZ, S. M., CURLESS, B., DIEBEL, J., SCHARSTEIN, D., AND SZELISKI, R., 2013. Middlebury multi-view stereo evaluation. http://vision.middlebury.edu/mview/.

SHALOM, S., SHAMIR, A., ZHANG, H., AND COHEN-OR, D. 2010. Cone carving for surface reconstruction. In *Proc. of SIGGRAPH Asia*, vol. 29, ACM, 150:1–150:10.

SHARF, A., LEWINER, T., SHAMIR, A., KOBBELT, L., AND COHEN–OR, D. 2006. Competing fronts for coarse–to–fine surface reconstruction. *Computer Graphics Forum (CGF) 25*, 3, 389–398.

SHEN, C., O'BRIEN, J. F., AND SHEWCHUK, J. R. 2004. Interpolating and approximating implicit surfaces from polygon soup. In *Proc. of SIGGRAPH*, ACM, New York, NY, USA, 896–904.

SINHA, S. N., MORDOHAI, P., AND POLLEFEYS, M. 2007. Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh. In *Proc. of International Conference on Computer Vision (ICCV)*, IEEE Computer Society.

SINHA, S. N., STEEDLY, D., AND SZELISKI, R. 2009. Piecewise planar stereo for image-based rendering. In *Proc. of International Conference on Computer Vision (ICCV)*, IEEE Computer Society, 1881–1888.

SNAVELY, N., SEITZ, S. M., AND SZELISKI, R. 2006. Photo tourism: Exploring image collections in 3D. In *Proc. of SIGGRAPH*, vol. 25, ACM, 835–846.

SNAVELY, N., SEITZ, S. M., AND SZELISKI, R. 2008. Skeletal sets for efficient structure from motion. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society.

SORMANN, M., ZACH, C., BAUER, J., KARNER, K., AND BISHOF, H. 2007. Watertight multi-view reconstruction based on volumetric graph-cuts. In *Image Analysis*, B. Ersbøll and K. Pedersen, Eds., vol. 4522 of *Lecture Notes in Computer Science (LNCS)*. Springer Berlin Heidelberg, 393–402.

STOLLNITZ, E. J., DEROSE, T. D., AND SALESIN, D. H. 1996. *Wavelets for Computer Graphics: Theory and Applications*. Morgan Kaufman.

STRECHA, C., VON HANSEN, W., GOOL, L. J. V., FUA, P., AND THOENNESSEN, U. 2008. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society.

SZELISKI, R. 2010. *Computer Vision: Algorithms and Applications*. Springer, New York.

TAGLIASACCHI, A., OLSON, M., ZHANG, H., HAMARNEH, G., AND COHEN-OR, D. 2011. VASE: Volume-aware surface evolution for surface reconstruction from incomplete point clouds. *Computer Graphics Forum (CGF) 30*, 5, 1563–1571.

TAYLOR, C. J. 2003. Surface reconstruction from feature based stereo. In *Proc. of International Conference on Computer Vision (ICCV)*, IEEE Computer Society, 184–190.

TRIGGS, B., MCLAUCHLAN, P. F., HARTLEY, R. I., AND FITZGIBBON, A. W. 2000. Bundle adjustment — a modern synthesis. In *Vision Algorithms: Theory and Practice*, B. Triggs, A. Zisserman, and R. Szeliski, Eds., vol. 1883 of *Lecture Notes in Computer Science (LNCS)*. Springer Berlin Heidelberg, 298–372.

TURK, G., AND LEVOY, M. 1994. Zippered polygon meshes from range images. In *Proc. of SIGGRAPH*, ACM, 311–318.

UNSER, M. 1997. Ten good reasons for using spline wavelets. In *SPIE Conference on Mathematical Imaging*, vol. 3169.

VOGIATZIS, G., HERNÁNDEZ ESTEBAN, C., TORR, P. H. S., AND CIPOLLA, R. 2007. Multi-view stereo via volumetric graph-cuts and occlusion robust photo-consistency. *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*.

WHITAKER, R. 1998. A level-set approach to 3D reconstruction from range data. *International Journal of Computer Vision (IJCV) 29*, 3, 203–231.

WILLIAMS, T. L. 1999. *The Optical Transfer Function of Imaging Systems*. Institute of Physics Publishing.

YANG, Q., YANG, R., DAVIS, J., AND NISTER, D. 2007. Spatial-depth super resolution for range images. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society.

YOON, K.-J., AND KWEON, I.-S. 2005. Locally adaptive support-weight approach for visual correspondence search. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society.

ZACH, C., POCK, T., AND BISCHOF, H. 2007. A globally optimal algorithm for robust TV-L1 range image integration. In *Proc. of International Conference on Computer Vision (ICCV)*, IEEE Computer Society.

ZHAO, H.-K., OSHER, S., AND FEDKIW, R. 2001. Fast surface reconstruction using the level set method. In *Proc. of Workshop on Variational and Level Set Methods in Computer Vision (VLSM)*, 194–201.

Curriculum Vitae

Personal Data

Name	Ronny Klowsky
Born in	Jena, Germany

Education

2008 – 2013	Technische Universität Darmstadt PhD student at Graphics, Capture and Massively Parallel Computing, Dept. Computer Science, under the supervision of Prof. Dr.-Ing. Michael Goesele
2002 – 2007	Friedrich-Schiller-Universität Jena Studies of Mathematics (Dipl.-Math.) Minor subject: Computer Science

Work Experience

2008 – 2013	Technische Universität Darmstadt Research Associate at Graphics, Capture and Massively Parallel Computing

Publications Co-Authored by R. Klowsky

The thesis is partially based on the following publications:

1. Ronny Klowsky, Michael Goesele. *Wavelet-based Surface Reconstruction from Multi-Scale Sample Points*. Technical Report 13rp006-GRIS, 2013.

2. Ronny Klowsky, Arjan Kuijper, Michael Goesele. *Weighted patch-based reconstruction: linking (multi-view) stereo to scale space*. In Proc. of Scale Space and Variational Methods in Computer Vision (SSVM), volume 7893 of LNCS, Schloss Seggau, Graz region, Austria, 2013.

3. Ronny Klowsky, Arjan Kuijper, Michael Goesele. *Modulation Transfer Function of Patch-based Stereo Systems*. In Proc. of IEEE Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 2012.

4. Ronny Klowsky, Patrick Mücke, Michael Goesele. *Hierarchical Surface Reconstruction from Multi-Resolution Point Samples*. In Post-Proceedings of the Dagstuhl-Workshop on Outdoor and Large-Scale Real-World Scene Analysis, volume 7474 of LNCS, 2012.

5. Patrick Mücke, Ronny Klowsky, Michael Goesele. *Surface Reconstruction from Multi-Resolution Sample Points*. In Proc. of Vision, Modeling, and Visualization (VMV), Berlin, Germany, 2011.

6. Mate Beljan, Ronny Klowsky, Michael Goesele. *A Multi-View Stereo Implementation on Massively Parallel Hardware*. Technical Report 08rp015-GRIS, 2011.

7. Michael Goesele, Jens Ackermann, Simon Fuhrmann, Carsten Haubold, Ronny Klowsky, Drew Steedly, Richard Szeliski. *Ambient Point Clouds for View Interpolation*. In Proc. of ACM SIGGRAPH 2010, Los Angeles, USA, 2010.

8. Michael Goesele, Jens Ackermann, Simon Fuhrmann, Ronny Klowsky, Fabian Langguth, Patrick Mücke, Martin Ritz. *Scene Reconstruction from Community Photo Collections*. In IEEE Computer (Issue June 2010), Invited Paper.

Advised Theses

1. Patrick Mücke. *3D Surface Reconstruction from Multi-Resolution Depth Maps*. Diploma thesis, 2011.

2. Sebastian Lipponer. *Point Based Rendering of Multi-View Stereo Depth Maps*. Diploma thesis, 2010.

3. Thorsten Franzel. *Verbesserte Structure-from-Motion Kalibrierung durch Bildverzerrung*. Bachelor thesis, 2009.

4. Mate Beljan. *Massively Parallel Implementation of a Multi-View Stereo Algorithm*. Bachelor thesis, 2008.