Andreas Schönfelder, BSc

Fusion of Large-Scale Aerial Imagery derived Point Clouds

MASTER'S THESIS

to achieve the university degree of

Diplom-Ingenieur

Master's degree programme: Geomatics Science

submitted to

Graz University of Technology

Supervisor

Univ.-Prof. Dr.rer.nat. Dipl.-Forstwirt Mathias Schardt

Institute of Geodesy

Graz, August 2017

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master‘s thesis.

Date

Signature

Abstract

Traditional methods in the field of photogrammetry and remote sensing mostly use aerial imagery for the generation of orthophotos and 2.5D digital surface models (DSMs), especially in rural areas, while light detection and ranging (LiDAR) was the predominant acquisition method for the reconstruction of dense cityscapes. Ongoing innovations in high resolution camera systems and the development of dense image matching algorithms have pushed the limits of image based 3D reconstruction regarding robustness, accuracy and performance. As a result, 3D reconstruction from aerial imagery has become a considerable alternative showing high potential for many applications. The objective of this thesis was the development of a scalable 3D point cloud fusion strategy, which builds on classic depth map based multi-view stereo (MVS) methods and fuses multiple stereo derived point clouds. The fusion of point clouds aims at the reduction of point cloud artifacts like outliers and noise while generating a non-redundant consistent surface representation. The method presented in this work enhances information obtained from stereo matching by computing a set of normal maps and classifying disparity maps in quality classes based on total variation (TV). With this information given, a local filtering strategy is applied. The strategy comprises the fusion of oriented point clouds along the surface normals, while prioritizing high quality disparities. The potential of the fusion method was evaluated based on airborne imagery (oblique and nadir) and satellite imagery. The results demonstrated that the fusion strategy is applicable for the generation of consistent surface representations of large-scale aerial imagery data sets consisting of billions of points.

Kurzfassung

Herkömmliche Methoden im Bereich der Photogrammetrie und Fernerkundung verwenden Luftbilder hauptsächlich zur Erstellung von Orthophotos sowie digitalen Oberflächenmodellen (DOM), insbesondere in ländlichen Gebieten, während Light Detection and Ranging (LiDAR) die bevorzugte Akquisitionsmethode für die Rekonstruktion von urbanen Gebieten darstellte. Innovationen im Bereich hochauflösender Kamerasysteme sowie die Entwicklung von neuen Bildkorrelationsverfahren (Dense Image Matching) haben das Potential automatisierter 3D Rekonstruktion hinsichtlich Robustheit, Genauigkeit und Leistungsfähigkeit gesteigert. Infolgedessen entwickelte sich die 3D Rekonstruktion aus Luftbildern zu einer potentiellen Alternative für eine Vielzahl an Anwendungen. Das Ziel dieser Arbeit war die Entwicklung einer skalierbaren Methode zur Fusionierung von aus Luftbildern abgeleiteten 3D Punktwolken, welche auf klassischen Multi-View Stereo Systemen aufbaut. Zweck der Fusionierung ist die Erstellung einer konsistenten 3D Oberflächenrepräsentation, welche frei von Beeinträchtigungen wie Ausreißern, Rauschen sowie Redundanz ist. Die in dieser Arbeit vorgestellte Methode nützt die im Zuge der 3D Rekonstruktion gewonnenen Informationen zur Ableitung von Oberflächennormalen sowie einer qualitätsbasierten Klassifizierung der berechneten Disparitäten. Basierend auf diesen Zusatzinformationen wird eine lokale Filterstrategie angewandt, welche die einzelnen Punkte entlang der zuvor berechneten Oberflächennormalen fusioniert und zugleich hochqualitative Disparitäten bevorzugt. Das Potential der vorgestellten Methode wurde anhand von flugzeug- sowie satellitengestützten Luftbilddatensätzen evaluiert. Die Ergebnisse zeigten, dass die vorgestellte Methode für die Generierung von konsistenten 3D Oberflächenmodellen aus großräumigen Luftbilddatensätze geeignet ist.

Acknowledgements

First of all, I would like to thank my supervisor, Univ.-Prof. Dr.rer.nat. Dipl.-Forstwirt Mathias Schardt for giving me the opportunity to work on this topic at the Joanneum Research within the working group Remote Sensing and Geoinformation.

Special thanks go to Karlheinz Gutjahr and Roland Perko for their support, advice and numerous discussions which helped me throughout this work.

Furthermore, I would like to thank Stefan Cavegn and Norbert Haala for providing the terrestrial laser scanner reference data, which helped me in the evaluation of this thesis.

Finally, I wish to thank my parents, family, and friends for supporting me throughout my life.

This research was partly done within the COMET programme, project nr. 836630, by Land Steiermark through SFG under project nr. 1000033937, and by the Vienna Business Agency.

Contents

1 Introduction

1.1 Motivation

Reconstructing three-dimensional (3D) models of real-world objects from imagery is and has always been a fundamental purpose of photogrammetry. By now, 3D reconstruction from imagery has also become a vivid research area in the field of computer vision with a large number of different applications, ranging from small-scale reconstruction utilizing optical microscopes up to large-scale aerial mapping applications with the aid of high-resolution airborne and satellite imagery.

Up to this time, aerial imagery was mostly used for the generation of orthophotos and 2.5D digital elevation models, especially in rural areas, while light detection and ranging (LiDAR) was the predominant acquisition method for the reconstruction of dense cityscapes. Ongoing innovations in high-resolution camera systems and the development of new multi-view stereo (MVS) algorithms have pushed the limits of automatic 3D reconstruction regarding robustness, accuracy and performance. As a result, 3D reconstruction from aerial imagery has become a considerable alternative showing high potential for many applications. Moreover, the availability of affordable unmanned aerial vehicle (UAV) has made data acquisition significantly easier and has further drawn attention to the topic.

In particular, the reconstruction from aerial imagery is challenging and per definition ill posed as large scenes have to be processed with, in most cases, limited computational capabilities. The large amounts of data can easily exceed the storage volume of the main memory and therefore need to be processed sequentially, which can lead to high runtimes. In order to overcome this problem, reconstruction algorithms have to scale well to the specific data set and ideally, should not be restricted to specific processing hardware like computer clusters. For this reason, many MVS systems that are specialized on the processing of aerial imagery are based on dense stereo algorithms. These algorithms usually yield one depth map or point cloud per stereo pair and can therefore easily be parallelized. However, the resulting advantage regarding scalability entails the need for a depth map or point cloud fusion strategy, which fuses the derived depth maps into one consistent surface representation. Moreover, the fusion procedure offers the possibility to detect and eliminate outliers and reduce multiple point redundancy while improving the overall accuracy of the surface.

While the fusion of depth maps for the generation of 2.5D digital surface models is a well-studied problem (Kuhn, 2014), the fusion of complex 3D scenes poses several new challenges. However, the generation of 3D surface models becomes inevitable when the representation in form of 2.5D digital surface models is insufficient. This, and the fact that 3D surface models are well suited for the purpose of visualization, serve as a motivation for this thesis.

1.2 Objective

The aim of this thesis is the development of a scalable 3D point cloud fusion strategy, which builds on classic depth map based multi-view stereo algorithms and fuses multiple stereo derived point clouds into one consistent surface representation.
Errors induced in the process of stereo matching and caused by insufficient modeling of the sensor imaging geometry lead directly to 3D reconstruction errors. Hence, every stereo derived point cloud is afflicted by a certain amount of point cloud artifacts like outliers and noise.
Contrary to approaches simply merging multiple point clouds into one surface representation, by combining several point clouds without further processing, the proposed method fuses and in this course filters, multiple point clouds generating one consistent non-redundant surface representation.
In detail, the information obtained from dense stereo algorithms is enhanced by computing a set of normal maps and a quality based classification of disparity maps. The classification of disparities enables the restriction, respectively weighting of every individual observation based on their expected spatial uncertainty. Consequently, outliers can be detected while noise and multiple point redundancy is reduced. The normal maps represent the orientation of the point clouds and are utilized to fuse the individual stereo derived point clouds successively along the direction of the surface. The resulting fused point cloud can directly be used to generate meshed surface models by utilizing surface reconstruction algorithms designed for the reconstruction from oriented point clouds.
The scalability of the fusion routine is ensured by the implementation of a point cloud tiling scheme. The tiling scheme enables the fusion of large-scale data sets without limitations regarding hardware and without compromising the overall performance.

1.3 Outline

Since the proposed method is based on the concept of classic multi-view stereo systems, a general overview on the topic of 3D reconstruction from multiple images and the underlying geometric basics is given in Chapter 2. Chapter 3 provides a review on state of the art multi-view stereo methods, with special focus on the task of depth map fusion and generation of consistent 3D surface models. Moreover, recent MVS methods are discussed, concerning their scalability and potential for the reconstruction of large-scale data sets (i.e. aerial imagery data sets). In Chapter 4 the developed point cloud fusion and filtering strategy is described in detail. Chapter 5 covers the analysis of the proposed fusion method. The potential of the fusion method is evaluated based on airborne (oblique and nadir) and satellite imagery. In case of the oblique imagery data set, reference data from terrestrial laser scanners is used to validate the results generated by the fusion routine. Finally, Chapter 6 summarizes the proposed fusion strategy and provides an outlook on possible further improvements.

Parts of this thesis have been previously published in:

A. Schönfelder et al. (2017). "Fusion of Point Clouds derived from Aerial Images." In: *OAGM and ARW Joint Workshop*. 2. Wien, Austria, pp. 139–144. DOI: 10.3217/978-3-85125-524-9

2 3D Reconstruction from Multiple Images

This thesis deals with the scalable reconstruction of large aerial imagery data sets by introducing a novel 3D point cloud fusion strategy. The corresponding point clouds, respectively depth maps, originate from a classic stereo reconstruction pipeline. This chapter presents an overview on the fundamental working principles of multi-view stereo (MVS) in Section 2.1 and the geometric basics of 3D reconstruction from multiple images in Section 2.2.

Generally, the reconstruction of a three-dimensional object is achieved by acquiring measurements, which describes the targets surface by a set of observations. The characteristic of those observations (e.g. distances, angles etc.) can vary based on the type of sensor that is used. The underlying sensor techniques can be categorized into two classes: active and passive sensors.
Active sensors emit signals into the direction of the target's surface and measure the reflected signal and its energy. For example, LiDAR systems measure the distances based on the time between the emission and detection of the signal.
In contrast to active sensors, passive sensors do not emit signals on their own, but receive and measure natural emissions produced by other sources of energy, like the sun. One example for passive sensors are cameras which are designed to measure natural energy in form of electromagnetic waves at specific frequencies. In the process of image acquisition, the 3D object is mapped into the two-dimensional (2D) domain (Vu, 2011). Acquiring data from multiple positions enables us to derive distances to the targets surface by triangulation. The underlying process is equivalent to the depth perception of human beings, also called stereopsis or stereovision. The real world is captured with the eyes, from two slightly different views. The brain matches the similarities between both views and conveys an impression of depth. MVS can be considered as the inverse process of capturing an object in a fixed scene (Vu, 2011). The goal is therefore to recover the 3D information from multiple two-dimensional images.

2.1 Overview on Multi-View Stereo

The following section gives an overview on the basic components and operating principles of MVS systems. According to Kuhn (2014) the problem of 3D reconstruction from multiple images can be divided into five sub-problems:

- Image Registration and Orientation
- Stereo Matching
- Depth Fusion
- Triangulation
- Optimization

2.1.1 Image Registration and Orientation

The objective of image registration and orientation is to determine the geometrical model (i.e. interior and exterior orientation) for the entire input imagery. The registration process requires the knowledge of the underlying camera model. Matching homologous points in overlapping areas allows the orientation, respectively its optimization, of multi-image networks in a least squares manner. The optimization is based on a robust bundle block adjustment and leads to an accurate estimation of the multi-view geometry. Depending on the photogrammetric application, parts of the geometrical model can be determined in advance. For example, professional aerial imagery is usually taken from large format mapping cameras whose interior orientation is known and where additional measurements like Global Positioning System (GPS) and inertial measurement unit (IMU) observations are available. However, for maximum accuracy it is also common to define the interior orientation parameters as observations of a specified accuracy rather than as constant. By doing so, their values are estimated during the bundle adjustment. This procedure is called self calibration and can be inevitable in cases like the registration of crowd-sourced amateur imagery where the parameters of interior orientation have to be estimated for every image individually (Schindler, 2014). As a byproduct, the image registration leads to a sparse 3D point cloud, which approximately represents the captured scene. The process of image registration and orientation equals the term structure from motion (SFM) as it successfully estimates camera poses (motion) and a sparse 3D representation (structure) (Kuhn, 2014).

2.1.2 Stereo Matching

In the process of stereo matching a depth map or disparity map is derived for every possible stereo constellation (i.e. stereo pair). The disparity is defined by the distance of two corresponding pixels within a stereo pair and comprises the 3D information of the respective object point (i.e. the stereo parallaxes in epipolar direction). In general, the disparity value is computed by measuring the similarity of pixels considering a specific type of pixel matching cost (Vu, 2011). The evaluation of the cost function (i.e. the computed matching cost for all potentially relevant disparities) yields the disparity value of the specific pixel.

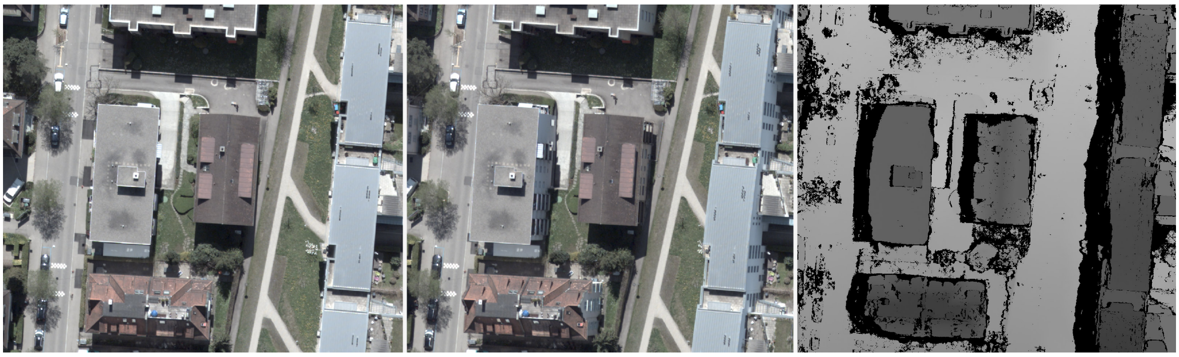


Figure 2.1: Stereo matching two epipolar rectified images resulting in a so-called disparity map *(right)* (i.e. the stereo parallaxes in epipolar direction).

The epipolar geometry, which describes the projective geometry between two views (see Section 2.2.2), can be used to generate epipolar rectified images in which the search for pixel correspondence is limited to one dimension (see Figure 2.1). The epipolar rectification must be performed for every stereo pair, prior to the actual image matching. In the simplest form of multi-view stereo, disparity maps are computed for every stereo pair independently. This leads to a number of redundant observations as every reference image is matched to a set of overlapping candidate images. In order to generate a consistent non-redundant surface representation the disparity maps or depth maps have to be fused (see Section 2.1.3). However, it is possible to improve the results obtained from stereo matching by linking the disparities from the rectified image to the input image. In this way, redundant observation can be utilized to check geometric consistency and to perform multi-view forward intersection. The multi-baseline triangulation yields a depth map per input image rather than per stereo pair (Rothermel, 2017).
Details on the the geometrical principles of 3D reconstruction from multiple images, including the epipolar geometry and epipolar rectification, are explained in Section 2.2)

2.1.3 Depth Map Fusion

Whereas many MVS methods directly generate consistent surface representation by linking points directly in the process of image matching and iteratively grow the final surface, depth map based methods require an additional fusion step. Dense depth maps or disparity maps which are produced in the course of image matching represent the scene with a certain degree of redundancy. The aim of depth map fusion is to merge the single sub-reconstructions into one consistent surface, while improving precision and reduce redundancy (Rothermel et al., 2016). Depending on the intended final surface representation (i.e. point clouds, 2.5D or 3D triangulated mesh etc.) different approaches have been developed. In Chapter 3 a review on state-of-the-art depth map and point cloud fusion algorithms with special focus on large-scale applications is given.

2.1.4 Triangulation and Optimization

Triangulation describes the transformation of point clouds to a connected set of polygons. This step may be considered as a post-processing procedure and is necessary if the depth map fusion yields point clouds, rather than a 3D model. As mentioned before, many MVS methods directly produce consistent surface representation in the course of image matching or depth map fusion, whereby triangulation becomes an inherent part of the MVS pipeline.

Optimization entails the modification of the final surface representation and can be performed as an optional post-processing step. For example, optimization in terms of mesh simplification is used to reduce the numbers of triangles, with the least possible loss in surface quality.

2.2 Geometric Basics

2.2.1 Single-View Geometry

As mentioned before, an image represent the projection of the 3D scene onto a 2D domain. The problem of obtaining 3D information from 2D images is inverse and hence, ill-posed. To make the reconstruction problem well posed additional knowledge is necessary. For every input image a corresponding camera model must be defined which fully describes the projection of a 3D point of the real world into a 2D pixel location of the image (Furukawa and Hernández, 2015). The most common camera model is the pinhole camera model which is also known as the central perspective projection. In order to fully describe the mapping with a perspective camera model, the parameters of the intrinsic and extrinsic orientation have to be defined.
The intrinsic orientation represents the projection from the camera coordinates to the image coordinates. The parameters are defined by the focal length c, the optical center (i.e. principal points) x_0 and y_0, as well as the geometric distortion induced by the lens of the camera, summarized by the terms Δx and Δy.
The extrinsic orientation is achieved by a transformation from the origin of the object coordinate system to the origin of the camera coordinates, followed by the rotation from the object coordinate system into the camera coordinate system (McGlone, 2013). Thus, the parameters of the extrinsic orientation can be described by the camera translation $\boldsymbol{X_0}$ and rotation $\boldsymbol{R}$.

The collinearity equation expresses the geometrical process of a central perspective projection. It states that the camera center $\boldsymbol{O'}$, the object point $\boldsymbol{P}$ and its corresponding image point $\boldsymbol{P'}$ lie on the same line. The coordinates of the object point $\boldsymbol{P}$ can therefore be derived from the vector to the camera center $\boldsymbol{X_0}$ and the vector from the camera center to the object point $\boldsymbol{X^*}$.

$$\boldsymbol{X} = \boldsymbol{X_0} + \boldsymbol{X^*} \tag{2.1}$$

However, the vector $\boldsymbol{X^*}$ can not be determined directly. Instead, the image vector $\boldsymbol{x'}$ can be transformed into the object coordinate systems by the rotation matrix $\boldsymbol{R}$ and the scale factor λ.

$$\boldsymbol{X^*} = \lambda \boldsymbol{R}\boldsymbol{x'} \tag{2.2}$$

Thus, the mapping of an image point into the object space is defined by

$$\begin{gathered}\boldsymbol{X} = \boldsymbol{X_0} + \lambda \boldsymbol{R}\boldsymbol{x'} \\ \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} X_0 \\ Y_0 \\ Z_0 \end{pmatrix} + \lambda \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix} \begin{pmatrix} x' - x'_0 - \Delta x' \\ y' - y'_0 - \Delta y' \\ -c \end{pmatrix}\end{gathered} \tag{2.3}$$

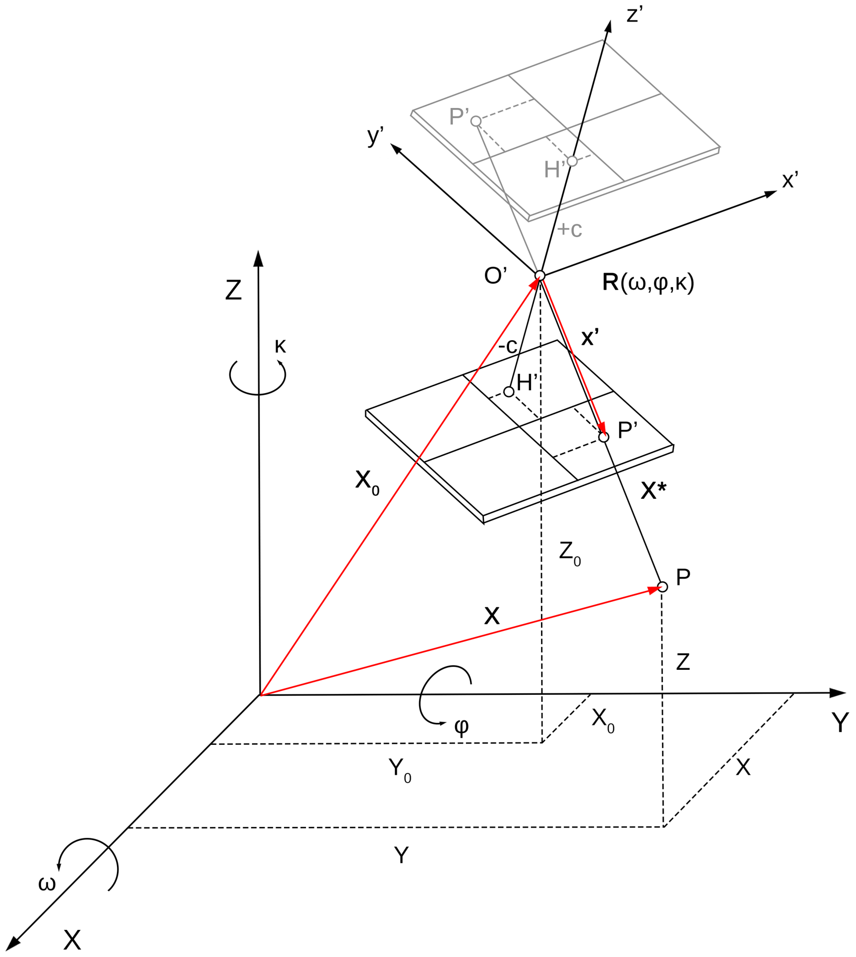


Figure 2.2: Extrinsic orientation and central perspective projection (based on Luhmann, 2010).

By dividing the equations defining the x and y components with the z component, the unknown scale factor λ is eliminated. Consequently, the collinearity equation can be written in the form

$$\begin{aligned} x' &= x'_0 - c\,\frac{r_{11}(X - X_0) + r_{12}(Y - Y_0) + r_{13}(Z - Z_0)}{r_{31}(X - X_0) + r_{32}(Y - Y_0) + r_{33}(Z - Z_0)} + \Delta x' \\ y' &= y'_0 - c\,\frac{r_{11}(X - X_0) + r_{22}(Y - Y_0) + r_{23}(Z - Z_0)}{r_{31}(X - X_0) + r_{32}(Y - Y_0) + r_{33}(Z - Z_0)} + \Delta y' \end{aligned} \tag{2.4}$$

where the image coordinates x and y (i.e. measured observations) are defined as a function of the unknown parameters (i.e intrinsic, extrinsic parameters and object point coordinates) (Luhmann, 2010). Whereas, depending on the photogrammetric problem, parts of the unknown parameters can be determined in advance.

Besides the pinhole camera model, optical line scanners are broadly used in the acquisition of remote sensing data. The imaging principle of these camera models is generally more complex. While a perspective relationships equivalent to the pinhole camera model is assumed in across-track direction, the movement of the sensor yields a time-varying component in along-track direction.

2.2.2 Two-View Geometry

Once the geometric model is defined, 3D point information can be obtained utilizing images taken from two different camera positions. A vector triangulation from two oriented views to the corresponding 2D points on the image planes leads to the coordinates of the 3D point. This relationship is expressed by the coplanarity constraint (or epipolar constraint), which states that the viewing rays of corresponding image points must lie on a plane since they intersect in one 3D point (see Figure 2.3). The intrinsic geometry, also known as the epipolar geometry, describes the relative orientation of the cameras. Even if only the relative orientation is known, it is possible to reconstruct a straight line-preserving, projectively distorted model of the object by intersecting the corresponding rays. If, additionally, the intrinsic orientation of the camera is determined, the reconstruction leads to an angle-preserving model. By adding extrinsic information of the camera positions the scale of the model is also fixed (Schindler, 2014).

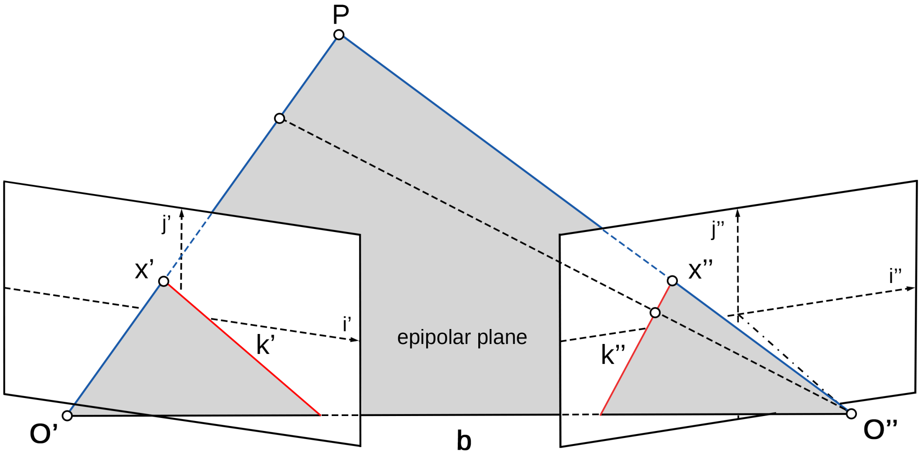


Figure 2.3: Epipolar geometry: Two views with their camera centers $\boldsymbol{O'}$ and $\boldsymbol{O''}$ observe an object point P. The projection of P onto the image planes leads to the corresponding image points x' and x'' in the camera coordinate systems (defined by i', j' and i'', j''). The viewing rays of the corresponding image points x' and x'' lie on the epipolar plane and intersect in P. The intersection of the epipolar plane with the image planes defines the epipolar lines k' and k'' (based on Luhmann, 2010).

The epipolar geometry is represented by a 3x3 matrix called the Fundamental matrix $\boldsymbol{F}$. If a point $\boldsymbol{P}$ in the real world is imaged as $\boldsymbol{x'}$ in view one and $\boldsymbol{x''}$ in view two, they satisfy the equation

$$x'^T F x'' = 0 \tag{2.5}$$

In the case of calibrated cameras (i.e. the intrinsic orientation is known) the relative orientation can be described by the essential matrix $\boldsymbol{E}$ which is a specialization of the fundamental matrix $\boldsymbol{F}$ (Hartley and Zisserman, 2004).

To reconstruct a dense model, corresponding image points have to be found for every single pixel. However, the estimation of corresponding points without prior information is complex and computationally expensive, particularly because of the large search area (Kuhn, 2014). The epipolar constraint states that a point defined in one view is represented by an epipolar line in the other view, on which the corresponding point lies (cf. Figure 2.3) (Hartley and Zisserman, 2004). Consequently, the search for corresponding image points can be limited to the epipolar line if the epipolar geometry is known. In practice, this step is realized by the generation of epipolar rectified images, also known as the so-called stereo normal case. As depicted in Figure 2.4, the images are resampled and the search area is reduced to one dimension, which minimizes the computational costs of image matching.

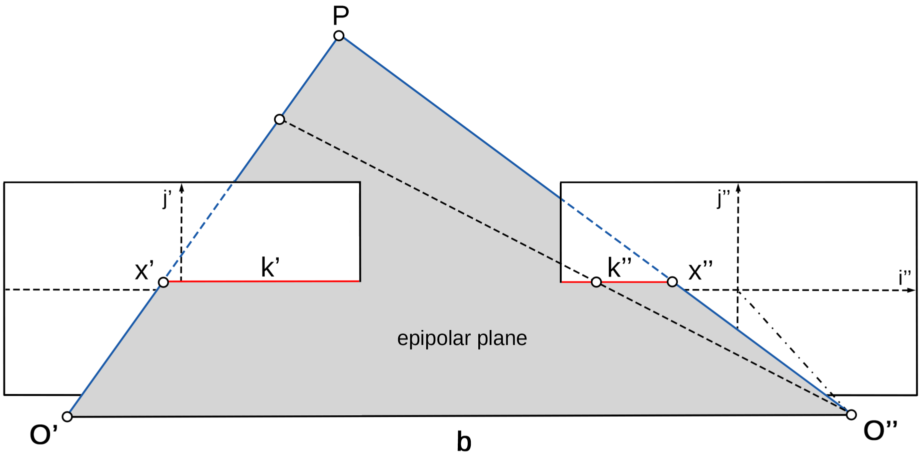


Figure 2.4: Epipolar geometry of the normal case: Reduction of the search area of corresponding image points to one dimension by the generation of epipolar rectified images (based on Luhmann, 2010).

For completeness, it shall be mentioned that beside the coplanarity in the case of two-view geometry, further constraints like the trifocal and quadrifocal constraint exists. These constraints describe the geometry for three and four image constellations. For a more detailed review on the mathematical foundations of multi-view geometry, reference is made to (Hartley and Zisserman, 2004; McGlone, 2013).

3 State of the art

This chapter presents a review of state of the art methods on the topic of Multi-View Stereo reconstruction, with focus on the generation of consistent 3D surface representations.

Over the past years, massive research regarding MVS systems has been conducted. The ongoing development of new MVS methods for different applications makes it hard to summarize all tendencies. Many algorithms are optimized to produce results under certain conditions with specific kinds of input data sets. This includes, for example, the requirement of lambertian surface or additional information such as silhouette information or semantic classification.
Seitz et al. (2006) proposed a MVS taxonomy which classifies methods based on six fundamental properties: scene representation, photo-consistency measure, visibility model, shape prior, reconstruction algorithm and initialization requirements. The work gives an overview on several MVS algorithms and serves as excellent reference for the topic. Furthermore Seitz et al. (2006) introduced the popular Middlebury MVS benchmark data set which has further driven the development of MVS methods. A different way to classify MVS methods is presented by Piracés (2008) where methods are divided into three groups: bottom-up, top-down and hybrid approaches. Whereas bottom-up techniques, directly extract 3D information from the images by matching patches between input images, top-down methods define an energy function relating the shape with the image information and then search for the minimum. Hybrid methods define a mixture of these techniques as they consist of an initial bottom-up step where 3D features are extracted and a subsequent top-down step where the surface is fitted to these features.
The reconstruction of large-scale data sets along with the processing of high-resolution imagery (i.e. aerial and satellite imagery) are vivid research areas, especially in the field of photogrammetry and remote sensing. While the processing of aerial and satellite imagery for the generation of 2.5D digital elevation models is a standard procedure (e.g. Perko et al., 2014) the reconstruction of 3D scenes poses several new challenges. Vu (2011) gives a detailed overview of recent MVS methods which are capable of reconstructing outdoor scenes, especially large-scale data sets.

At the beginning of this chapter, a brief introduction to different scene representations and their relationship to consistent surface generation is given.
Since the algorithm presented in this work builds on standard MVS pipelines and

assumes a set of computed depth maps as input data, consistent surface generation methods in terms of depth map fusion strategies are discussed in detail. Recent work concerning MVS concepts like photo-consistency measures, visibility models or shape priors are not part of this thesis an therefore not discussed in detail. The reader interested in this topic is referred to Seitz et al. (2006), Vu (2011), and Furukawa and Hernández (2015) as they summarize state-of-the-art MVS principles and concepts.
Finally, important work that is concerned with the reconstruction of large-scale data sets is discussed in Section 3.1.3.

3.1 Consistent 3D Surface Generation

Since the proposed algorithm deals with the topic of point cloud fusion, in this section recent work and different approaches for the generation of consistent 3D surface models from multiple images are discussed. The problem of consistent surface generation strongly depends on the concept of the underlying MVS method and scene representation. For example, some methods directly link points in the matching process to grow a consistent surface mesh, while others require an additional fusion step. Therefore, a classification of MVS methods based on the geometrical representation of reconstructed scene is given in the following section.

3.1.1 Scene Representation

Generally, the scene representation describes the data structure or mathematical framework of the extracted scene geometry. According to Seitz et al. (2006), these representations can differ from voxels, level-sets, polygon meshes up to depth maps. Depending on the reconstruction algorithm, different representations are employed in various steps of the reconstruction pipeline.
Voxels represent the geometry on a three dimensional grid by adding occupancy information. The simplicity, uniformity and ability to approximate any surface make 3D grids (i.e. volumes) a popular representation for different MVS techniques, such as Visual Hull (Baumgart, 1974), Voxel Coloring (Seitz and Dyer, 1999) and Space Carving (Kutulakos and Seitz, 1999).
Polygon meshes are another popular output format for MVS methods as they are efficient to store and render (Seitz et al., 2006). They represent the scene geometry by a set of connected planar facets and define a consistent 3D surface model where redundancy is usually eliminated in the reconstruction pipeline. This is possible by linking surface points directly in the process of image matching. For example, Furukawa and Ponce (2010) use multi-photo consistency measures to optimize positions and normals of surface patches and iteratively grow the surface starting from a set of feature points.

Other methods describe the scene by a set of depth maps. The representation with depth maps avoids resampling the geometry on a 3D domain (Seitz et al., 2006). However, to extract a consistent, non-redundant 3D surface representation (e.g. a polygon mesh) the depth maps have to be fused (see Section 3.1.2). Since dense stereo or multi-baseline MVS algorithms directly produce disparity or depth maps, this type of scene representation is broadly used. Moreover, depth map based MVS methods scale well to large data sets as the reconstruction problem can easily be split into a number of sub-problems.
Beside the mentioned scene representations, elevation maps, also referred to as digital surface models (DSMs) or digital terrain models (DTMs) are broadly used in the photogrammetric community. According to Seitz et al. (2006) elevation maps can be considered as an alternative representation of depth maps, as their depth values are defined relative to the scene's surface. This scene representation is suitable whenever it is sufficient to represent the scene's geometry by a 2.5D structure, for example for surface reliefs (Vogiatzis et al., 2008) or aerial imagery. However, instead of directly generating elevation maps in the reconstruction pipeline as presented in Pierrot-Deseilligny and Paparoditis, 2006, 2.5D digital elevation models are often produced in a subsequent set by employing a fusion strategy.

3.1.2 Depth Map Fusion

Depth map fusion or integration describes the process of merging 2.5D sub-reconstructions (i.e. depth maps) of the captured scene into one consistent non-redundant surface representation. Since the processing of multiple stereo images yield one depth map or disparity map per stereo pair, redundant information is generated.
The combination of several depth maps into one scene representation can be performed by using general surface from point cloud algorithms like (Hoppe et al., 1992; Amenta et al., 2001). However using these generic methods usually does not lead to satisfactory results, since the individual depth maps are distorted by point cloud artifacts like outliers and noise (Zach et al., 2007). Hence, the objective of depth map fusion is not only to generate a consistent surface representation but also to remove outliers, reduce redundancy and improve the geometric precision of the 3D reconstruction.
The fusion or integration of depth maps and point clouds has been an active research topic for decades primarily in the computer vision and the graphics community. However, the past development of dense image matching algorithms, like the Semi-Global Matching (SGM) (Hirschmüller, 2008), led to increased research activities on the topic, particular for large-scale photogrammetric applications.
In general, the problem of 3D surface reconstruction from a set of redundant depth estimations can be classified into two groups: either a local cost function, which considers a limited part of the data, or a global cost function over all the data is minimized (Kuhn, 2014). The distinction is important as the Middlebury benchmark

(Seitz et al., 2006) has already shown that global methods tend to produce the best results regarding completeness and accuracy, while local methods offer good scalability at smaller computational costs and the possibility of parallel computing.

Polygon techniques, like the Polygon Zippering proposed by Turk and Levoy (1994), triangulate depth maps in image space and lift the triangulation to the object space by utilizing the spatial information of the depth maps. The triangle meshes are generated locally by constructing faces from adjacent depth estimations and subsequently removing of suspicious or redundant faces. Remaining meshes are stitched together to connect boundaries and to generate the continuous surface representation.

As shown in Labatut et al. (2009), the problem of 3D reconstruction from multiple depth maps can also be defined as an energy minimization problem solved by a s-t cut optimization framework. In a first step, a Delaunay tetrahedralization of the depth samples and camera position is computed in object space. By constructing a directed graph the problem is solved by finding the s-t cutset of the graph, minimizing visibility information (i.e. line of sight intersecting the surface) and the quality of the reconstructed surface in terms of triangle size.

Many depth fusion algorithms build on volumetric range integration of depth maps, also called volumetric range image processing (VRIP). As stated in Zach et al. (2007), the use of intermediate volumetric surface representations allows the generation of models with arbitrary genus and avoids the numerical difficulties encountered with polygonal techniques. Early work based on VRIP, like Curless and Levoy (1996), compute 3D distance fields on an octree data structure by projecting depth estimation. The 2.5D range images are merged by employing an averaging sheme on the distance field. The final surface representation can then be obtained by minimizing an underlying energy function and using the Marching Cube method to triangulate the iso-surface (Lorensen and Cline, 1987).

Volumetric methods for 3D reconstruction were first developed for the fusion of range images acquired with laser scanners and have later been adapted for the fusion of disparity maps or depth maps. However, these methods can have high computational cost and large memory requirements, particularly for large-scale data sets, because the number of elements rises with the third power (Kuhn, 2014). Hilton and Illingworth (1997) have shown that volumetric methods are suitable for the reconstruction of large scenes by using octrees at different levels. Fuhrmann and Goesele (2011) expanded the use of (multi-level) octrees by storing vertices at different scales depending on the depth maps pixel footprint and considering the geometric uncertainty of 3D points. In this way a hierarchical distance field is generated from which the final surface representation is extracted by applying the Marching Tetrahedra algorithm (Akio and Koide, 1991). Kuhn et al. (2013) adapted the concept of multi-resolution voxels with dynamic sizes and shows that local volumetric methods are not limited concerning scalability when dividing the reconstruction space in independent subsets.

3.1.3 Scalable 3D Surface Generation

Especially for the processing of aerial imagery, scalability is an important factor. As mentioned in Kuhn et al., 2016b, a number of scalable fusion methods have been presented in the last years, e.g. (Fuhrmann and Goesele, 2011; Kuhn et al., 2013; Ummenhofer and Brox, 2015; Rothermel et al., 2016), yet they are still not able to process billions of 3D points in a single day or less (Ummenhofer and Brox, 2015). Kuhn et al. (2016a) propose a fast fusion method via occupancy grids for semantic classification. The fusion method complements state-of-the-art depth map fusion, as it is much faster. However, it is only suitable for applications that have no need for dense point clouds.

Kuhn (2014) proposes a scalable 3D surface reconstruction by local stochastic fusion of disparity maps. He shows that the 3D reconstruction of fused disparity maps can be improved by modeling the uncertainties of disparity maps. These uncertainties are modeled by introducing a feature based on total variation (TV) which allows pixel-wise classification of disparities into different error classes. Total variation in context with MVS was first introduced by Zach et al. (2007). They propose a novel range integration method using a global energy functional containing a TV regularization force and an L^1 data fidelity term for increased robustness to outliers. The minimization of the TV-L1 based global energy function yields the surface represented by a level set. Although this global volumetric method produces impressive results and offers the possibility of parallel execution on GPUs, the time and memory demands are significant which is why the methods is restricted to small and compact objects.

Authors like Galliani et al. (2015) and Rothermel et al. (2016) focus on the fusion of depth maps to generate oriented 3D point clouds. The surface reconstruction in terms of fitting a surface to the reconstructed and fused points is defined as a post-processing step, which can be solved using algorithms like the generic Poisson surface reconstruction method proposed by Kazhdan and Hoppe (2013). The Poisson surface reconstruction is an algorithm producing watertight meshes of excellent quality, taking oriented point clouds as input. Although both methods are defined as local in terms of depth map fusion and offer scalability, the surface reconstruction is global, as Poisson surface reconstruction is used for this step.

Rothermel et al. (2016) introduce a local median-based fusion scheme, which is robust to outliers and produces surfaces comparable to the results of the Middlebury MVS. Similar to Fuhrmann and Goesele (2011) points are subsampled using a (multi-level) octree. Favoring points with the smallest pixel footprint, an initial point set is created utilizing nearest neighbor queries optimized for cylindrical neighborhoods. Points are then filtered iteratively along line of sight or surface normals. The capability of the fusion strategy for large-scale city reconstruction and the straightforward manner for implementation make it particularly interesting for this work.

4 Development of a Point Cloud Fusion and Filtering Method

In this chapter, the developed point cloud fusion routine for the generation of consistent 3D surface models is presented. Chapter 2 presented the fundamental principles and geometric basics of 3D reconstruction from multiple images. Based on the treated basics and the review of state-of-the-art methods for large-scaleMVS reconstruction (see Chapter 3), an algorithm has been developed. The following chapter provides detailed information on the implementation of the algorithm.
Before explaining details of the fusion routine, an overview of the overall workflow is given. The fusion method builds directly on classic depth map based Multi-View Stereo reconstruction methods, which deal with multiple images as input data. For this reason, a rough overview of the underlying reconstruction pipeline is given in Section 4.1. Section 4.2 and 4.3 focus on pre-processing steps, which are part of the fusion pipeline and produce required input data. In Section 4.4 the employed point cloud tiling scheme, which ensures the scalability of the fusion method, is explained. Finally, a detailed explanation of the implemented fusion procedure is provided in Section 4.5.

4.1 Overall Workflow

The reconstruction pipeline covers the generation of consistent 3D surface models from a set of aerial input images. The 3D reconstruction (i.e. image orientation, stereo matching) leads to different intermediate results which are utilized in the fusion pipeline. Figure 4.1 depicts the overall workflow of the processing pipeline and illustrates the relation between the existing reconstruction procedure and the proposed fusion method.

4.1.1 3D Reconstruction

The proposed framework builds upon the Remote Sensing Package Graz (RSG)[1]. RSG may be considered as a toolbox for the processing of digital multi-sensor remote

[1] http://www.remotesensing.at/en/remote-sensing-software.html

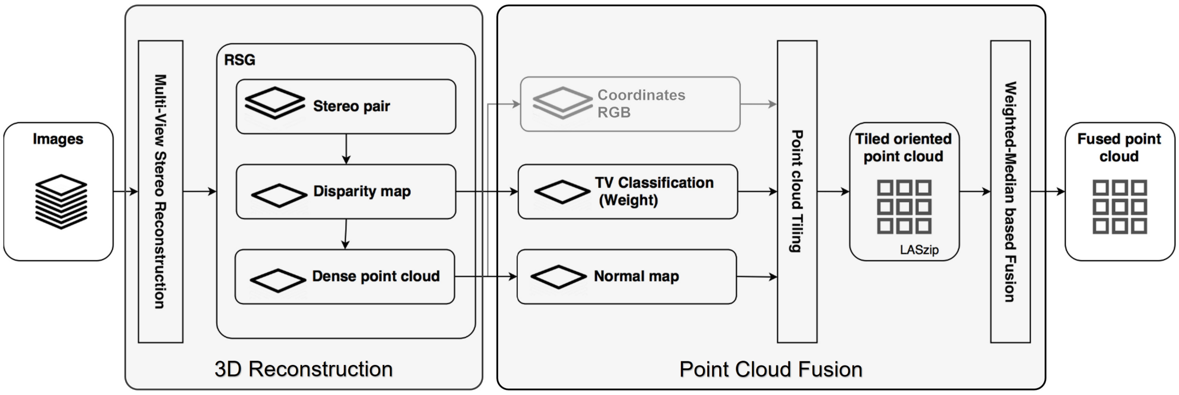


Figure 4.1: Workflow of the processing pipeline for point cloud fusion.

sensing data and offers, beside many different functionalities like synthetic aperture radar (SAR) interferometry, a full 3D reconstruction pipeline for remote sensing image data. However, considering the fusion of depth maps, the capability of RSG is currently limited to the generation of 2.5D digital elevation models. The proposed framework takes intermediate results of the 3D reconstruction pipeline as input parameters and enables the generation of consistent 3D surface representations.
In RSG disparity maps are derived from a set of epipolar rectified images using a matching algorithm based on SGM. After the matching procedure, forward and backward matching are employed to derive two dense point clouds per stereo pair. The coordinates are estimated via spatial point intersection and stored in East-North-Height (ENH) raster files (i.e. a three band raster file holding the coordinates in geometry of the disparity map). The advantage of this approach is that coordinates can be accessed directly while the spatial organization of the point cloud is preserved (i.e. the structure of the point cloud).

4.1.2 Point Cloud Fusion

As depicted in Figure 4.1 the disparity maps as well as the dense point cloud serve as input for the fusion procedure. A set of surface normals is calculated and stored in a so-called normal map. Moreover, the quality of the disparity maps is assessed to derive a quality based classification of the disparity map. A weight value is calculated for every class, which corresponds to the spatial uncertainty of disparity observations. These additional information is combined with the results of the 3D reconstruction pipeline (i.e. colorized dense point cloud). Subsequently, the point clouds are assigned to tiles in order to enable a tile-wise fusion of the data. Finally, the fusion procedure is employed resulting in one fused oriented point cloud per tile. The generation of meshed 3D surface models is defined as a post-processing step and therefore not an integral part of the processing pipeline (cf. Figure 4.1).

4.2 Oriented Point Cloud Generation

At the beginning of the point cloud fusion procedure, oriented point clouds (i.e. normal maps) are derived from the dense point cloud. Similar to Rothermel et al. (2016), the surface normals are utilized to iteratively filter single point observations. While in Rothermel et al. (2016) normals are derived based on a restricted quadtree triangulation (Pajarola, 1998), in this work, the surface normals are estimated in a least squares manner. In a first step, the local point neighborhood is extracted by applying a moving window operation on the ENH raster files. The ENH raster files holds the coordinates of the reconstructed point cloud (i.e. east, north and height components) in the geometry of the disparity map. Therefore, the surface normals can be computed by locally fitting a plane to the extracted point neighborhood (i.e. the extracted point cloud) (see Figure 4.2). The size of the moving window and therefore the size of the extracted neighborhood affects the smoothness of the calculated surface normals. The normal estimation fails in areas with less than three reconstructed disparities. By introducing a threshold defining a minimum number of successfully reconstructed points in the local neighborhood, the robustness of the normal calculation can be controlled.

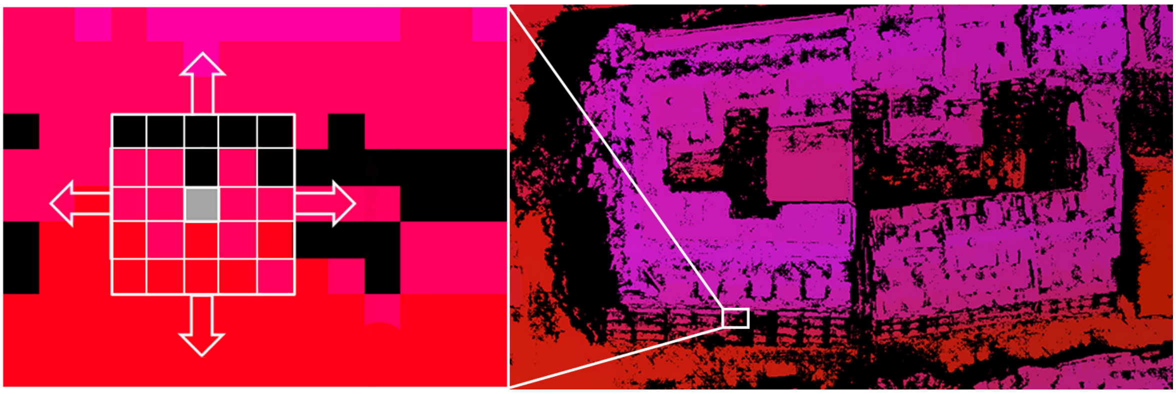


Figure 4.2: Extraction of local neighborhood *(left)* from ENH raster file *(right)*.

To improve the quality of the oriented point cloud further, visibilty checks are applied based on the acquisition geometry of the input images. As depcited in Figure 4.3, flipped and invalid surface normals are corrected or eliminated, based on the direction of the surface normal and camera orientations. Invalid surface normals are detected by analysing the range of the valid surface tilt angle. The basic principle of stereo reconstruction builds on the assumption that pixel correspondence between two views can be found (i.e. the surface has to be oriented towards both cameras). Hence, invalid surface normals defining a surface facing the cameras from different directions are eliminated.

The normal maps are calculated for every point cloud independently and are later combined with the existing information obtained from the 3D reconstruction (see Section 4.4).

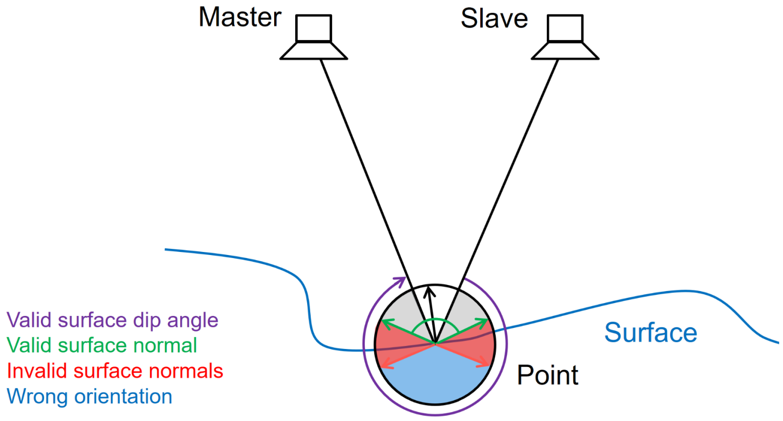


Figure 4.3: Analysis of surface normal orientation based on acquisition geometry.

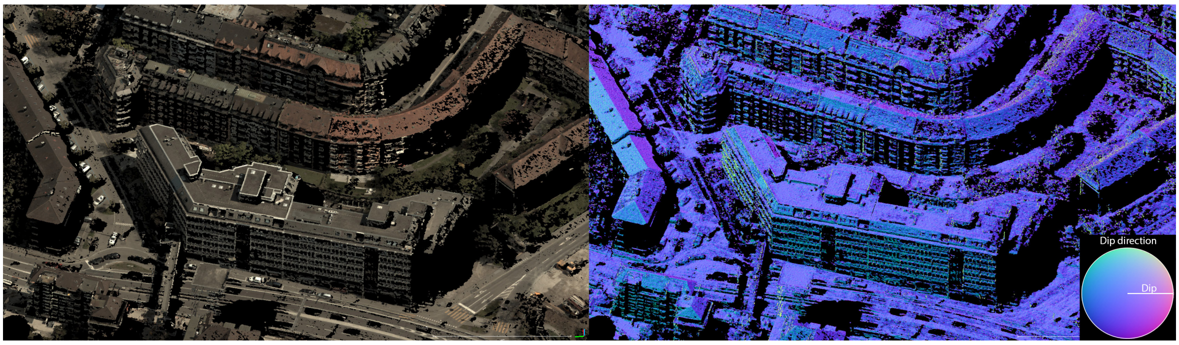


Figure 4.4: Colorized point cloud *(left)* and oriented point cloud with color-coded normals *(right)* derived from two oblique aerial images .

4.3 Disparity Quality Assessment

The following section focuses on the assessment of the disparity quality and its significance for local MVS methods.

4.3.1 Uncertainties in Stereo Matching

In general, 3D reconstruction errors are caused directly by errors of the disparities result. By reviewing the stereo error model proposed by Molton and Brady (2000) the disparity error Δp can be defined. Kuhn (2014) states that the ellipsoidal error model introduced by Molton and Brady (2000), propagates the disparity error Δp in image space into the error in object space $\Delta P_x, \Delta P_y$ and ΔP_z (Figure 4.5). While the error in x and y direction of the camera coordinate system rises linearly with P_z the error in z direction rises quadratically.

$$\Delta P_x = \Delta p \frac{P_z}{ft} \sqrt{(t - P_x)^2 + P_x^2} \tag{4.1}$$

$$\Delta P_y = \Delta p \frac{P_z}{ft} \sqrt{2P_y^2 + \frac{t^2}{2}} \tag{4.2}$$

$$\Delta P_z = \Delta p \frac{P_z^2}{ft} \sqrt{2} \tag{4.3}$$

The coordinates of the reconstructed point P_x, P_y and P_z are derived from the estimated disparities. The focal length f and baseline t are known for rectified images. The matched image coordinates p_1 and p_2, representing the object point $\boldsymbol{P}$ are distorted by the disparity error Δp, which is not known and not constant. Therefore, an uncertainty function, ranging from subpixel up to several pixels has to be considered. Especially for local MVS methods, it is important to consider this function, as these methods are not able to regularize uncertainties globally (Kuhn et al., 2016b). For more details on the applied stereo error model and the propagation of the disparity uncertainty, the reader is referred to Kuhn (2014).

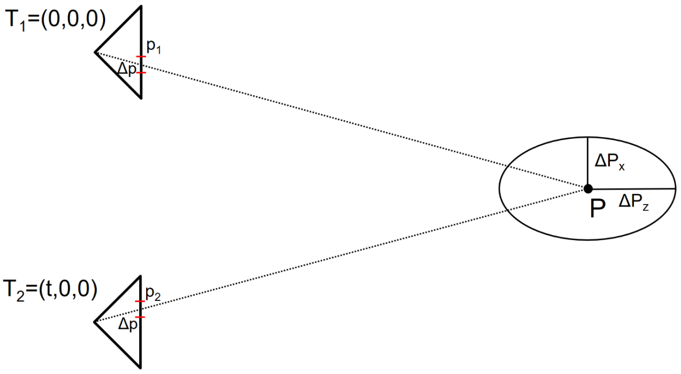


Figure 4.5: Ellipsoidal error model. The disparity uncertainty Δp resulting in an ellipsoidal uncertainty ΔP of the reconstructed 3D point P (based on Kuhn, 2014).

In this work, the errors introduced in the course of stereo matching by the disparities uncertainty are taken into consideration to improve the result of fusion routine. Unfortunately, there are several features, which influence the accuracy of disparity measurements, from which two would be texture strength and surface slant. Most stereo methods, including SGM, employ priors which favor constant disparities and therefore cause a fronto-parallel bias. Disparities are propagated from well textured into textureless regions and hence lead to errors on slanted and curved surfaces. As stated by Kuhn (2014), naively learning the qualities would lead to a multivariate system where the corresponding multivariate uncertainty has to be learned for all camera types and perhaps even all types of scenes. Therefore, Kuhn (2014) estimates the uncertainty from the disparity map directly by analyzing the local oscillation behavior of the disparity map.

4.3.2 Local Total Variation based Classification

For the proposed fusion method, the classification of disparity uncertainty introduced by Kuhn (2014) is adapted to derive weights, which are later used in the fusion procedure. The following section covers details of the local total variation (TV) based classification.

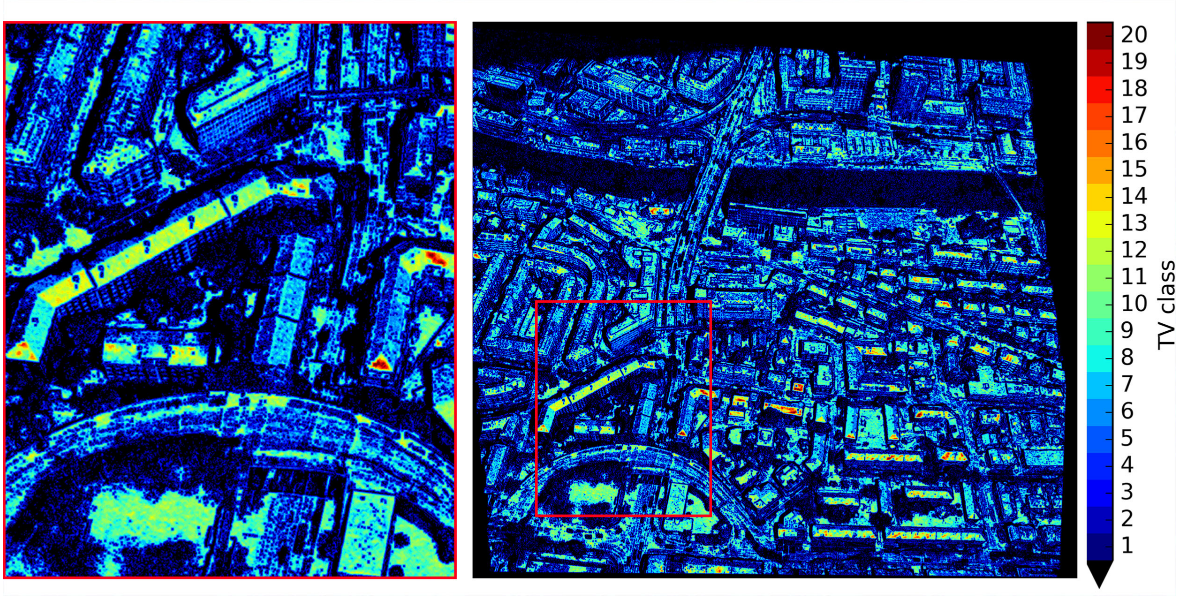
The main objective of the local TV is to measure the uncertainty of disparities. However, measuring the uncertainty is not trivial as disparities oscillate with unkown frequency. Moreover, learning the distribution of a 2D function can result to the estimation of wrong correlation. Kuhn (2014) deals with these problems by measuring the local oscillation using feature classes based on TV utilizing the L^2 norm (see Eq. 4.4). In contrast to many TV-L^1 based MVS methods, the L^2 norm takes noise and outliers into consideration, which is required to measure the quality of the disparities.

$$TV(y) = \sum_{i,j \epsilon \mathcal{N}_y} \sqrt{\left|d_{i+1,j} - d_{i,j}\right|^2 + \left|d_{i,j+1} - d_{i,j}\right|^2} \tag{4.4}$$

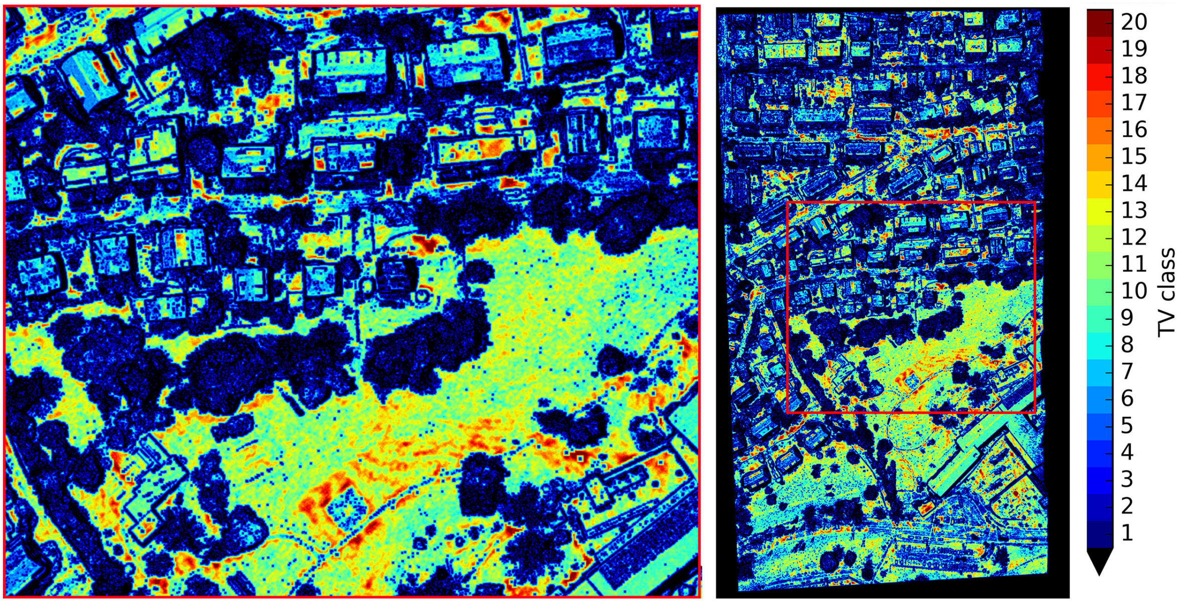
The TV is calculated over square windows $\mathcal{N}_y$ with increasing radius resulting in $n = [1, 20] \in \mathbb{N}$ discrete classes. By increasing the window size it is possible to measure the varying frequencies of the oscillating disparities. The discretization is achieved by introducing a regularization term τ which limits the TV to stay below a certain value (see Eq. 4.5). These TV classes describe the degree of the local oscillation of the disparities.

$$\arg\max_{n} = \left(\sum_{m=1}^{n} = \frac{1}{8m} TV_{i,j \in x_m} < \tau \right) \tag{4.5}$$

Figure 4.6 shows the TV classification for disparity maps derived from oblique and aerial images. As already mentioned, the disparities oscillation in a local neighborhood have different frequencies. Fronto-parallel planes cause low frequencies and therefore lead to higher TV classes, while slanted surfaces lead to higher frequencies and cause low classification values. The disparity map derived from two nadir aerial images, depicted in Figure 4.6b, shows a larger presence of high TV classes, due to the increased percentage of fronto-parallel surfaces (cf. Figure 4.6a).



(a) Disparity map derived from two oblique aerial images.



(b) Disparity map derived from two nadir aerial images.

Figure 4.6: Assessment of disparity quality based on local TV using a regularization term of $\tau = 2$.

4.3.3 Weight Function

The TV classification provides a discretization of the disparities uncertainty into different error classes. However, since the aim is to derive weights which represent the quality of the measured disparity, the TV classes are analyzed based on their geometric precision and accuracy.
As stated by Kuhn (2014), the outlier probability of disparity measurements can be

obtained by learning error distributions from this classification using ground truth disparities. In this work, error distributions are not learned directly, as there are no ground truth disparities available. Therefore, the quality of the classified disparities is analyzed in object space rather than image space. Reference data from terrestrial laser scanners (TLSs) is used to assess the quality of the raw dense point cloud for every single TV class independently.

In a first step, vertical DSMs are computed for facade patches where reference data is available, following the approach of Cavegn et al. (2014). Subsequently, the weights are computed in form of a weighting function, by analyzing the DSMs derived from the classified point cloud and the reference data for various test areas. The weighting function is estimated by calculating the standard deviation of the flatness error and fitting an exponential function in a least squares manner. The flatness error is defined as the point cloud deviations to a best fitting plane and is also an indicator for the noise of the 3D geometry (Ahmadabadian et al., 2013). The results of the estimated weight function and its effect considering the geometric precision (i.e. level of noise) are discussed in Section 5.

4.4 Point Cloud Tiling

Aerial imagery data sets are usually acquired by high-resolution camera systems and with high overlap to guarantee good coverage of the scene. The problem is that these data sets may reach critical sizes exceeding the capacity of the main memory. Therefore, MVS methods for large-scale reconstruction have to provide some sort of scalability.

One way to enable scalability of the reconstruction pipeline are the so-called Divide and Conquer methods. Divide and Conquer methods divide the 3D space into a number of subspaces. Individual parts are then processed independently allowing parallelization of the reconstruction problem.

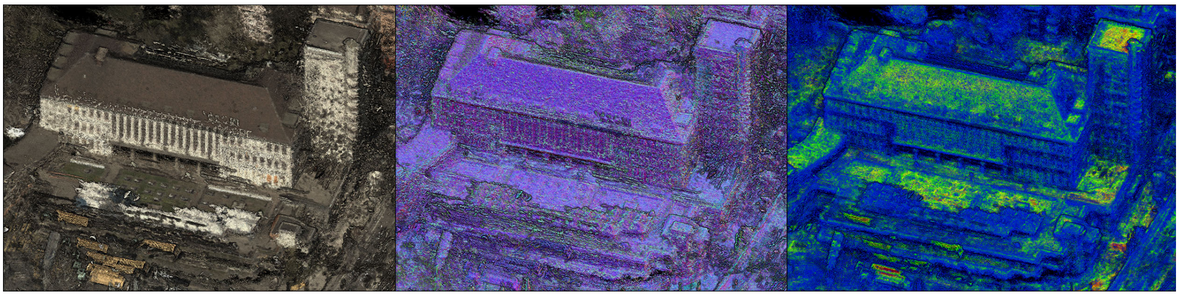


Figure 4.7: Combined and tiled point cloud generated from several stereo pairs containing a high ratio of outliers and noise. Colorized dense point cloud *(left)*, oriented point cloud with color-coded normals *(middle)* and classified point cloud based on local TV *(right)*.

In the proposed framework, the scalability is realized by following the Divide and Conquer approach. In a first step, the information for every stereo pair is combined (cf. Figure 4.1) and stored in a compressed LAS file (i.e. a lossless compressed data format for point cloud data) (Isenburg, 2013). Subsequently, the tiling functionality of the LASTools[2] software suite is employed. This function enables the tiling of large point clouds stored in compressed LAS files. The point clouds are assigned to square, non-overlapping tiles of a specified size. To avoid discontinuities on tile borders, a buffer is defined which grows the tile (temporarily) in every direction (see Figure 4.8). The employed tiling scheme solely allows the tiling in a 2.5D manner (i.e. no tiling is applied in Z direction). Nevertheless, this function is sufficient for the processing of aerial imagery. The tiled point cloud then serves as input for the fusion procedure, which is performed for every tile independently. Finally, the fused point clouds are put back together by utilizing the reversible tiling functionality of the LASTools software suite.

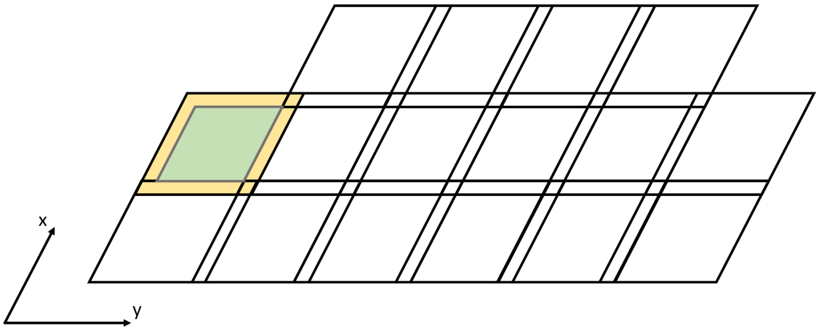


Figure 4.8: Point cloud tiling scheme.

4.5 Weighted-Median Based Fusion

The concept of median-based fusion originates from fusion algorithms for the generation of 2.5D DSMs. Rothermel et al. (2016) adapt this idea by fusing point clouds in 3D space along a defined filtering direction. While for close range data sets the line of sight is suitable as filtering direction, point-wise normals are used for the fusion of aerial data sets. In this work, the fusion strategy is adapted by using a weighted-median based approach utilizing weights derived from the disparity quality assessment. The fusion procedure yields one fused and oriented point cloud per tile. This chapter covers relevant parts of the fusion strategy introduced by Rothermel et al. (2016) and novel adaptations made in the course of this thesis. Figure 4.9 depicts the workflow of the proposed fusion routine.

Initial Pointset: In a first step, an initial pointset P is created from the input point cloud by storing the input point cloud in an octree data structure. The pointset P is derived by subsampling the point cloud with the centroid of the points located

[2] https://rapidlasso.com/lastools/

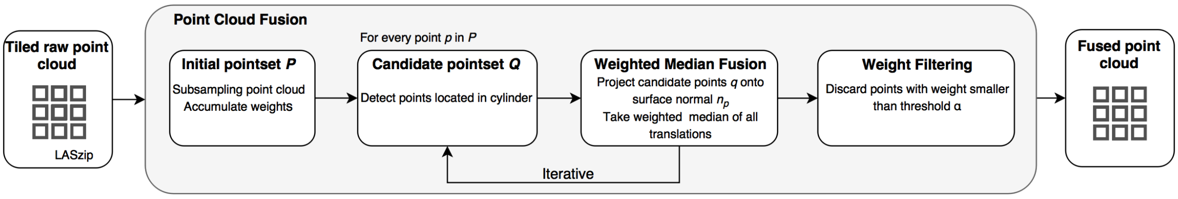


Figure 4.9: Workflow of the proposed fusion procedure.

in a leaf node. Positions and normal directions are averaged and stored in the initial pointset P. In order to reduce memory requirements Rothermel et al. (2016) implemented a hash map based octree. In this work, the entire fusion process was realized with the aid of the Point Cloud Library (PCL ver. 1.8.0) (Rusu and Cousins, 2011) which also provides a custom tailored octree implementation.
As a result of the disparity quality assessment, every point possesses a weight w_i representing the quality of the point. We add up the weights of all points located in the same leaf node. Thus, the weight of the initial point $p \in P$ is an indicator for the density and quality of the reconstructed scene.

Candidate Pointset: Once the initial pointset P is defined, the point cloud is fused. For every point in the initial pointset P a set of candidate points Q, located in a cylinder with its central axis given by the initial point $\boldsymbol{p}$, its normal $\boldsymbol{n_p} = (n_x, n_y, n_z)$, a height h and radius r, is derived (see Figure 4.10). The identification of this pointset is performed using nearest neighbor queries optimized for cylindrical neighborhoods, introduced by Rothermel et al. (2016).
Starting from the mother nodes octree box, each internal node is checked if itself or any child node contains leaf node points located in the cylinder. To detect if a candidate octree box B contains any leaf points located in the cylinder, the center $\boldsymbol{c}$ of the octree box B is transformed into the coordinate system defined by the initial point $\boldsymbol{p}$ and its normal $\boldsymbol{n_p}$.

$$\boldsymbol{c}' = \boldsymbol{R}(\boldsymbol{c} - \boldsymbol{p}) \tag{4.6}$$

The axis of the coordinate systems are given by the columns of the rotation matrix $\boldsymbol{R}$.

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{r}_1^\top \\ \boldsymbol{r}_2^\top \\ \boldsymbol{r}_3^\top \end{pmatrix} = \begin{pmatrix} 1 & 1 & -\frac{n_x + n_y}{n_z} \\ & (\boldsymbol{r}_1^\top \times \boldsymbol{r}_3^\top)^\top & \\ n_x & n_y & n_z \end{pmatrix} \tag{4.7}$$

The octree Box B may contain points located in the cylinder if the following two conditions are met. First, the horizontal distance of the transformed box center $\boldsymbol{c}'$ to the coordinate origin is compared to the radius of the cylinder (see Eq. 4.8).

The radius is increased by the term $\sqrt{3}s$, whereas the term $\sqrt{3}s$ corresponds to the radius of a sphere enclosing the octree box B. In a second step, the height of the cylinder is checked.

$$\sqrt{(c'_x)^2 + (c'_y)^2} < r + \sqrt{3}s \tag{4.8}$$

$$c'_z < h \tag{4.9}$$

If conditions 4.8 and 4.9 are not fulfilled the traversal of child nodes is terminated. However, if both conditions are met and the candidate box B is a leaf node, all points located in B are a subset of the candidate pointset Q.
After all potential candidate points $q \in Q$ are detected, the locations of these points are verified. Points that are not located in the specific cylinder are removed from the pointset Q. Following the Eq. 4.6 - 4.9, points are checked by exchanging the box center c with the point location $\boldsymbol{q}$.
Additionally, the angle between the surface normals $\boldsymbol{n_p}$ and $\boldsymbol{n_q}$ are computed. Points with surface normals diverging by more than 60° are discarded to avoid the incorporation of points representing different surfaces (Rothermel et al., 2016).

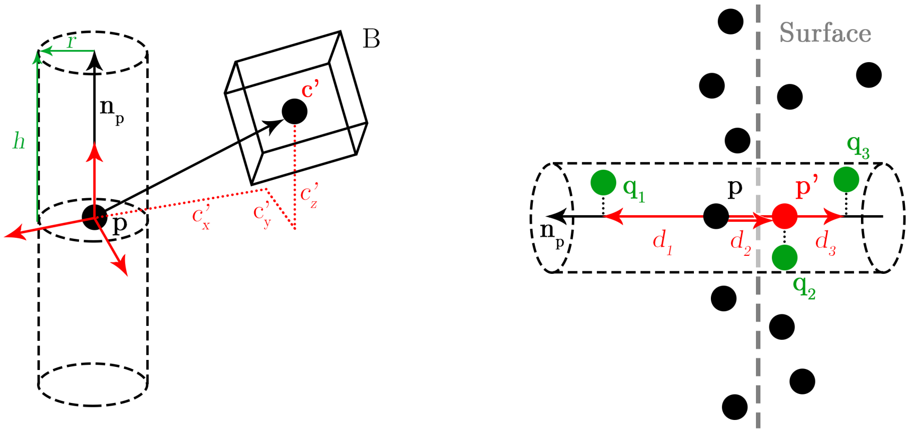


Figure 4.10: Identification of candidate pointset Q *(left)*: Evaluation of conditions describing if an octree cube B contains points inside a cylinder defined by $\boldsymbol{p}, \boldsymbol{n_p}, h, r$. The octree center $\boldsymbol{c'}$ is analyzed with respect to a coordinate system defined by the cylinder.
Weighted-median based filtering *(right)*: After all candidate points located in the cylinder are detected, the coordinates of the filtered point $\boldsymbol{p'}$ are derived by projecting the candidate points $\boldsymbol{q} \in Q$ onto the surface normal $\boldsymbol{n_p}$ and taking the weighted-median of the translations d_i (based on Rothermel et al., 2016).

Weighted-Median Fusion: After the candidate pointset Q is detected, the point $\boldsymbol{p}$ is filtered by projecting all candidate points $q \in Q$ onto the surface normal of the initial point $\boldsymbol{p}$ (see Figure 4.10). Taking the weighted-median of all deviations d_i to the point $\boldsymbol{p}$ yields the new point coordinates $\boldsymbol{p'}$.

$$d_i(\boldsymbol{q_i}) = (\boldsymbol{q_i} - \boldsymbol{p}^{\intercal}\boldsymbol{n_p}) \tag{4.10}$$

$$\boldsymbol{p'} = \boldsymbol{p} + \boldsymbol{n_p}\text{weighted-median}(d_i(\boldsymbol{q_i}), w_i) \tag{4.11}$$

Especially for noisy data, further iterations are inevitable to generate a consistent surface representation. Between every iteration, duplicate points, with respect to a minimal distance between points, are united to avoid redundant computations.

In a first iteration, Rothermel et al. (2016) includes all points of the input point cloud for the identification of the candidate pointset Q. To speed up further iterations, the filtering is restricted to the initial pointset P solely. In this work, filtering of the point cloud is restricted to the initial pointset P from the beginning on. The loss of detail, caused by subsampling the input cloud (i.e. creation of the initial pointset) is compensated by approximating the density of the captured 3D scene with the accumulated weight. However, in this way it is possible to fuse large and highly redundant 3D point clouds in moderate time (e.g. processing 2.5 billion points on a computer with 8 cores and hyper-threading within a single day, resulting in a fused point cloud whose density fits the spatial resolution of the input imagery).

Weight Filtering: As mentioned before, the accumulated weight is an indicator for the density and quality of the reconstructed scene. Points showing little support in terms of a small accumulated weight value are more likely to represent outliers. Therefore, the result obtained from the weighted-median based fusion can be improved by discarding points with weights smaller than a defined threshold α (see Figure 4.11). The influence of parameter values and the potential of the fusion routine are analyzed in Section 5.

Surface Reconstruction: Similar to Galliani et al. (2015) and Rothermel et al. (2016), the surface reconstruction in terms of fitting a surface to the reconstructed point cloud is solved by utilizing the Screened Poisson Surface Reconstruction[3] algorithm proposed by Kazhdan and Hoppe (2013). The fused and filtered oriented point cloud is directly used as input parameter for the reconstruction method (cf. Figure 4.11).

[3] http://www.cs.jhu.edu/~misha/Code/PoissonRecon/Version9.01/

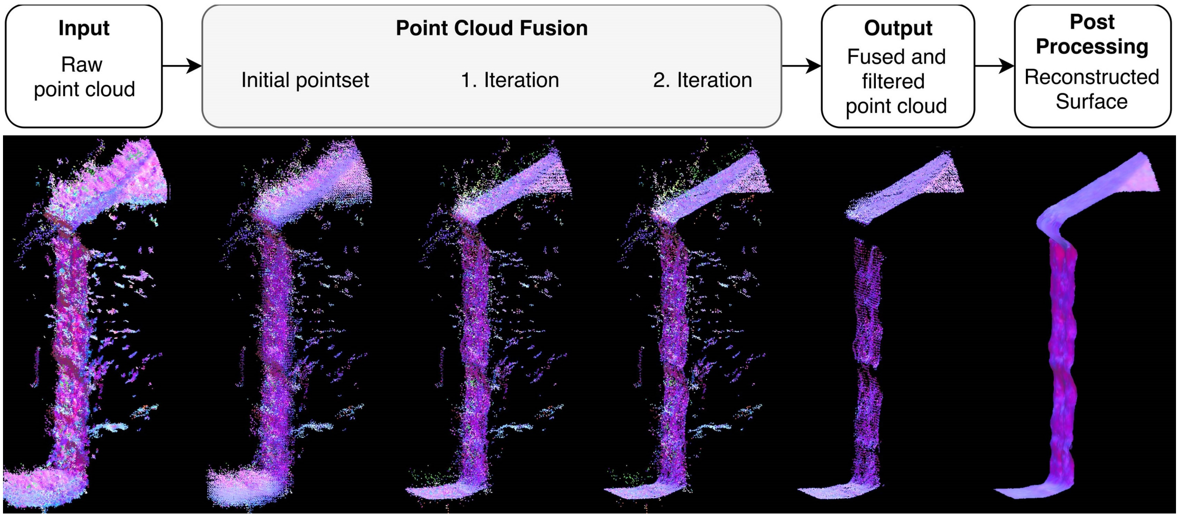


Figure 4.11: Intermediate results of the proposed fusion procedure. Taking the raw point cloud as input data and, after the initial pointset is derived, fuses the points successively along the surface normals. The final surface representation is derived by discarding points with a weight lower than a certain threshold. The fused and filtered oriented point cloud can then be meshed in a post-processing step.

5 Results and Evaluation

The following chapter presents analyses and results of the proposed fusion method. The potential of the fusion procedure is evaluated based on airborne (oblique and nadir) and satellite imagery.
The results obtained from the oblique imagery data set are validated with the aid of reference data acquired by two terrestrial laser scanners. A comparison of the raw and fused point clouds, with respect to the ground truth data, allows to measure the potential of the fusion routine and its capability to improve the quality of the reconstructed scene. In case of the nadir aerial and satellite imagery data sets no reference data is available. Hence, the evaluation is restricted to visual assessment of the output data and analysis of the fusion routines performance.

5.1 Oblique Aerial Imagery

Test Data: The airborne data sets, both nadir and oblique, are provided by the ISPRS/EuroSDR project on "Benchmark on High Density Aerial Image Matching"[1]. The project aims at the evaluation of photogrammetric 3D reconstruction in view of the ongoing developments of software for automatic image matching.
Based on the oblique imagery data set, the potential of the 3D reconstruction pipeline and the proposed fusion routine for the reconstruction of 3D city models is assessed. Airborne oblique views depict, in contrast to nadir views, facades and other vertical structures. Hence, it is possible to reconstruct 3D city models including vertical information. However, the processing of oblique views introduces a number of new challenges. For example, multiple occlusions caused by urban structures or greater changes in viewpoints and direction resulting in increased differences in object scale and larger variations in radiometry (see Figure 5.1).

The oblique imagery data set was acquired over the city of Zürich with a Leica RCD30 Oblique Penta medium format camera. The Leica RCD30 Oblique Penta camera is a professional mapping camera and consists of five 80 megapixel (MP) camera heads oriented in a maltese cross configuration, with one nadir view and four oblique views inclined by an angle of 35°. The nadir imagery was acquired with an overlap of 70% along flight and 50% across flight direction. While the nadir

[1] `http://www.ifp.uni-stuttgart.de/ISPRS-EuroSDR/ImageMatching/`

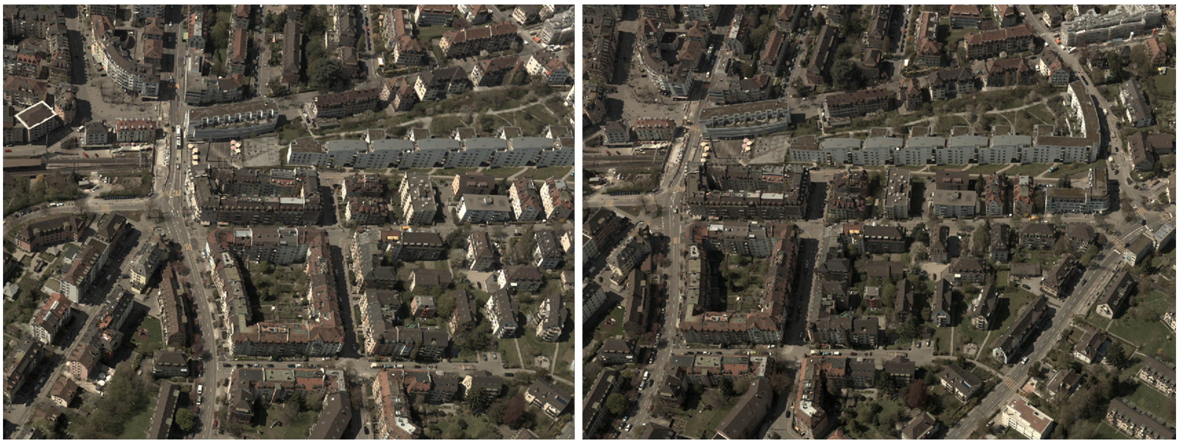


Figure 5.1: Subset of stereo configuration comprising two oblique views.

imagery is captured at a ground sampling distance (GSD) of 6 cm, the GSD of the oblique views vary between 6 and 13 cm. The entire data set consists of 27 unique camera positions, resulting in 135 images and covering an area of approximately 1.5 x 1.5 kilometers (see Figure 5.2).

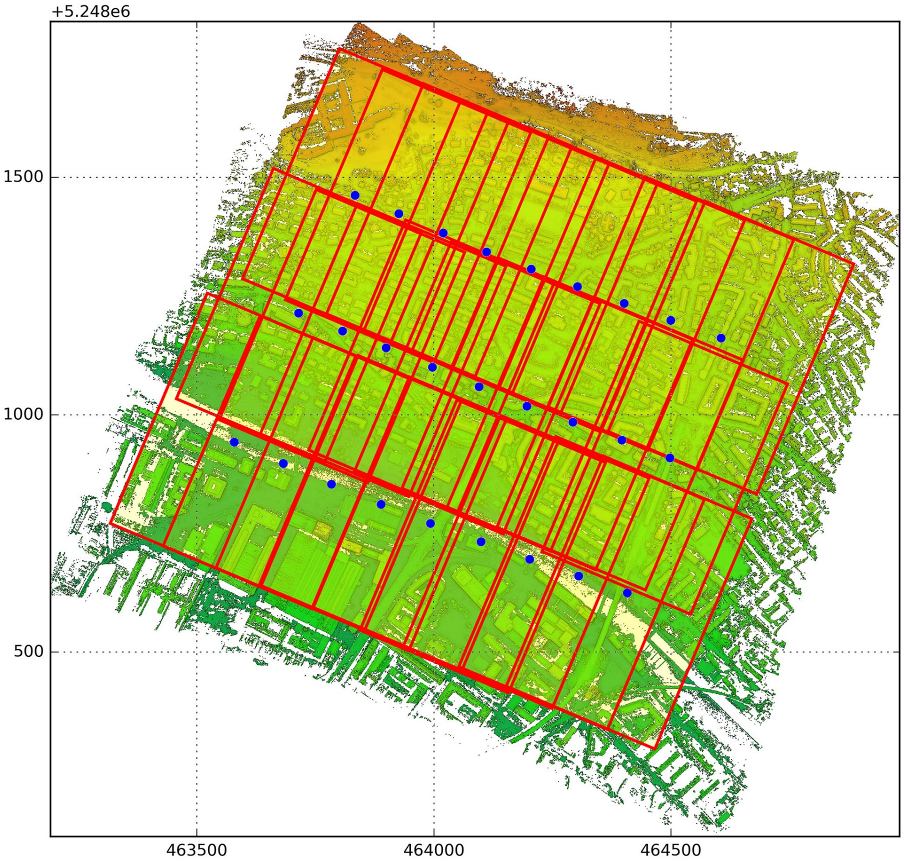


Figure 5.2: Oblique imagery: Camera positions and nadir image footprints overlaid to the reconstructed scene.

Reference Data: Reference data captured with TLSs provide accurate and reliable information for the evaluation of the data set. The laser scans were acquired by a Leica ScanStation 2 and a Leica ScanStation P20. For each scanner position, four to five points were measured in point cloud (RTK) mode to guarantee accurate registration of the laser scans. According to Deuber (2014) and Cavegn et al. (2014), the mean absolute accuracy of the provided registered TLS points lies within 2.2 cm, which meets the 3D accuracy expectation of 1/3-1/2 of the GSD.

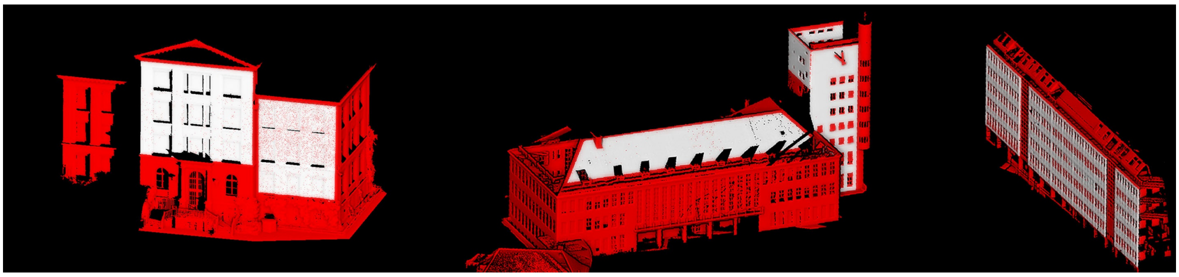


Figure 5.3: Reference TLS point clouds and extracted facade patches of two selected test areas: school building *(left)* and tower area with neighboring building *(right)*.

Figure 5.3 shows two selected test areas and extracted facade patches which were used for the verification of the fusion method. More information on the image acquisition, benchmark and reference data can be found in Cavegn et al. (2014).

5.1.1 3D Reconstruction

The following Sections cover details on the employed processing pipeline, particularly for the specific data set, by describing selected parameters, intermediate results, problems occurred within the processing of the data, as well as an analysis of the final results. For an overview on the overall workflow and its implementation, reference is made to Section 4.1.

In a first step, the 3D reconstruction was performed with the aid of the Remote Sensing Package Graz (RSG). The image orientation and registration was carried out using the interior and exterior orientation parameters provided along with the image data. Subsequently, images were matched in flight direction with an overlap of 70% resulting in a total of 314 stereo-pairs. Forward and backward matching yield 628 disparity maps and ENH raster files, containing approximately 10.6 billion points. In total, the resulting ENH raster files require up to 680 gigabyte to store. Due to the high overlap and good coverage of the scene the reconstructed point clouds are highly redundant. Moreover, the point clouds are afflicted by outliers and noise, leading to distortions of up to several meters (see Figure 5.4). It is worth mentioning that, in some cases, during the image acquisition parts of the helicopter skids protruded into the camera angle, which led to further errors in the matching.

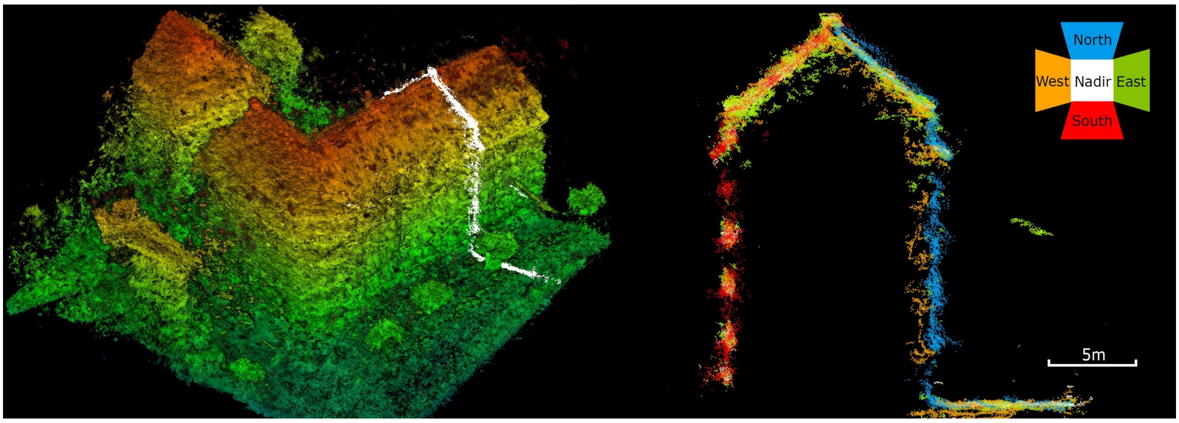


Figure 5.4: Subset of raw reconstructed point cloud combined from several stereo pairs with colorized height component *(left)*. Cross profile with a width of 10 centimeters *(right)*. Points color-coded based on their originating camera.

5.1.2 Pre-processing

TV classes and normal maps are computed for every stereo pair independently and serve as input for the fusion routine. The weighting function assigns a weight to every TV class which is then used in the fusion process. Following the approach introduced in Section 4.3.3, the weighting function was derived by analyzing DSMs for various test areas. Figure 5.5 illustrates the correlation between TV classes and outlier probability in 3D space.

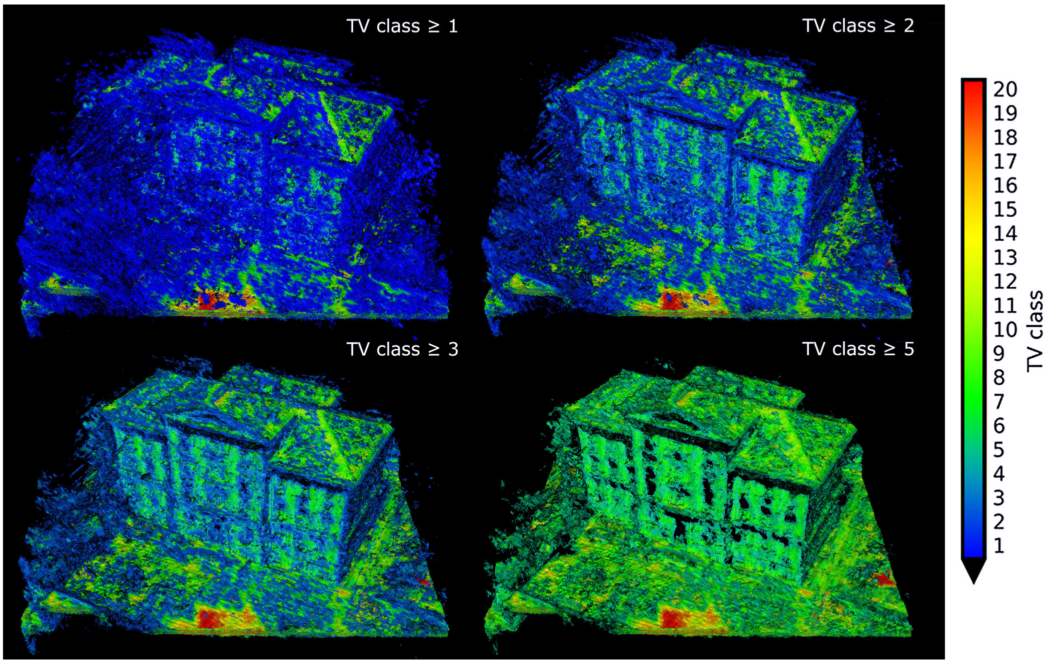


Figure 5.5: Raw dense point cloud restricted to different TV classes.

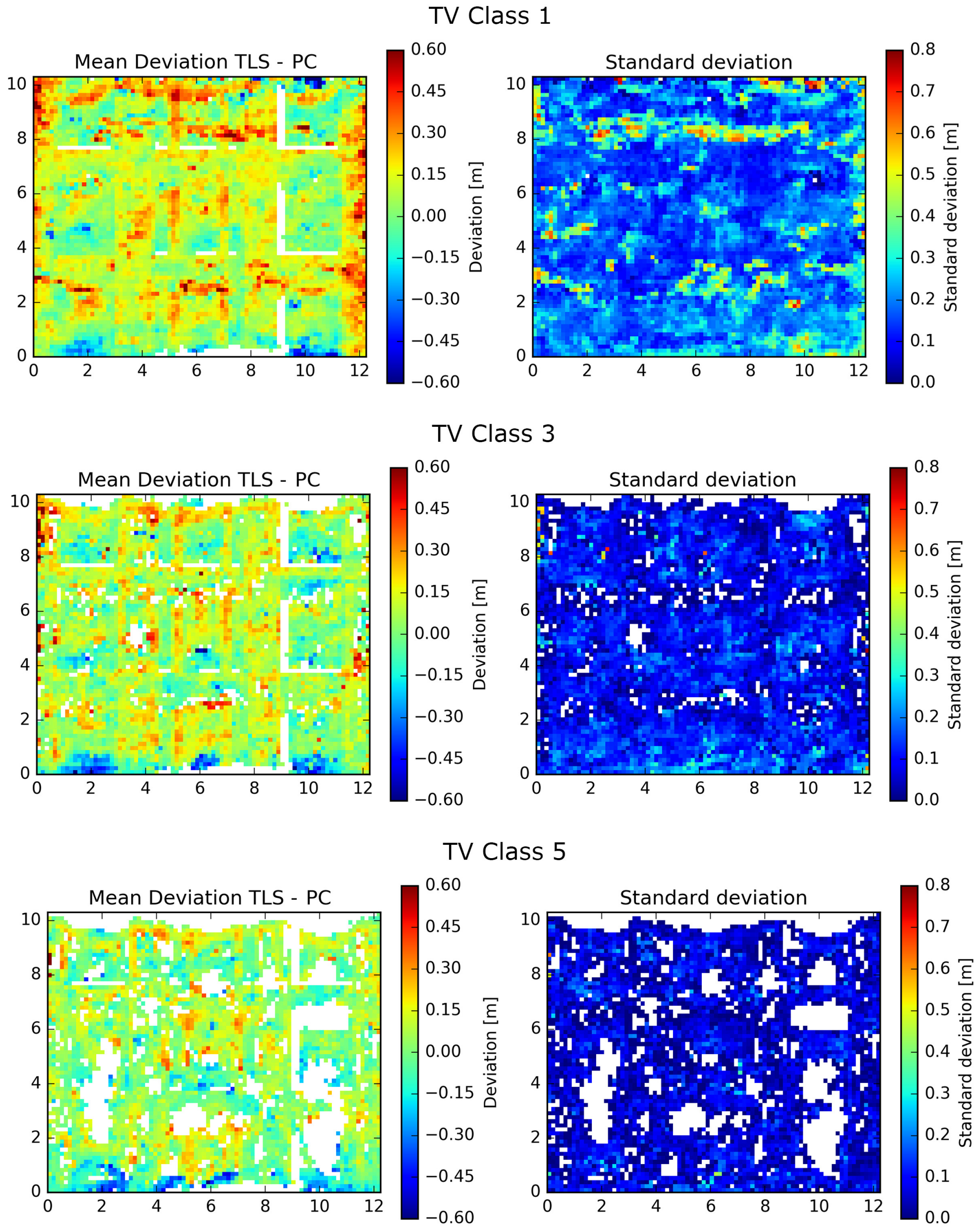


Figure 5.6: Analysis of TV classes based on vertical DSMs. DSMs computed for front facade of the school building test area.

Figure 5.6 illustrates the mean deviation of the point cloud with respect to the reference TLS data and the standard deviation of the point cloud (PC) for a subset of TV classes. The derived weighting function is depicted in Figure 5.7 and shows that a correlation between TV classes and the geometric precision (i.e. level of noise) can be verified. While higher TV-classes show smaller standard deviations and deliver better overall accuracy, lower TV-classes are more likely to contain outliers (also cf. Figure 5.5, Figure 5.6). TV classes greater than 8 are only present in flat areas facing the camera position. Since this thesis focuses on the reconstruction of vertical surfaces (i.e. facades) the information obtained by the test areas was extrapolated for all TV classes. Finally, the weighting function was derived by inverting the estimated function and defining the minimum weight with 1.0.

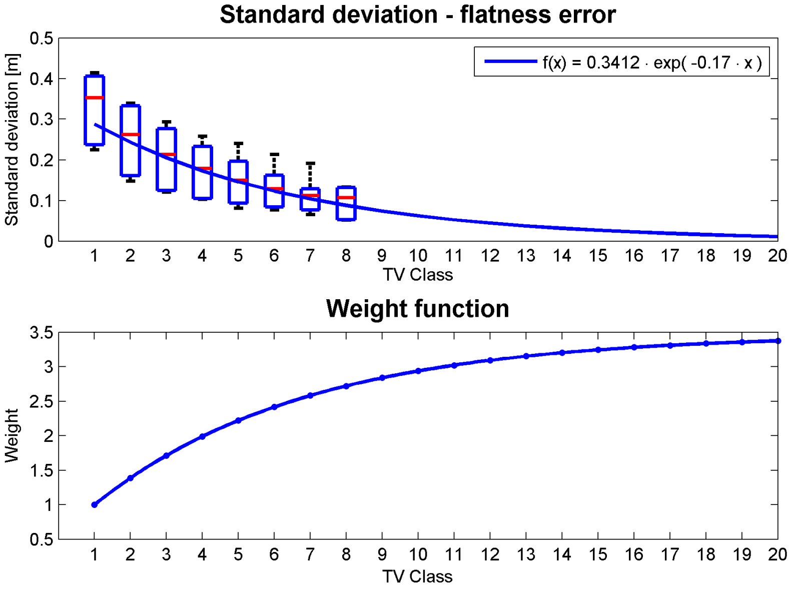


Figure 5.7: Box plots representing the standard deviation of the flatness error derived from different test areas for all available TV classes *(top)*. Estimated weight function *(bottom)*.

5.1.3 Point Cloud Fusion

The fusion of the point cloud was carried out in three iterations with a cylinder radius of 15 cm (i.e. approx. two times the GSD) and a total height of 3.0 m. The size of the octrees leaf node (i.e. the voxel resolution), which is used for the generation of the initial pointset, controls the approximate output density of the fused point cloud. Therefore, faster runtimes can be achieved by producing point clouds with lower density. The runtime of the fusion process can also be improved by discarding low-level TV classes in a pre-processing step. However, the rejection of low-level TV classes causes a loss in detail in areas with bad coverage.
The voxel resolution used for the oblique imagery was set to 10 cm to match the GSD of the input data. Within the point cloud fusion process, the points are filtered along the surface normal and weights are accumulated. The final surface representation is derived by discarding low weights, which are more likely to contain outliers. As depicted in Figure 5.8, increasing the minimum weight threshold α leads to more accurate, however less dense, point clouds.

Intermediate results of the fusion routine are depicted in Figure 5.9. Figure 5.10 shows the fused point cloud for a larger area. Since the fusion method produces oriented point clouds, a mesh representation was computed using the Poisson surface reconstruction algorithm (Kazhdan and Hoppe, 2013). Figure 5.11 illustrates the mesh generated from tiles of the fused point cloud and the potential of the

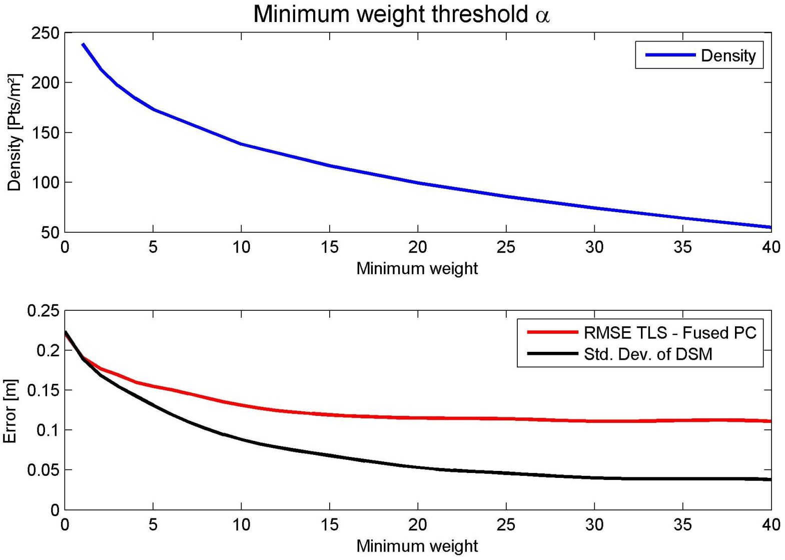


Figure 5.8: Impact of rejecting low weighted points after the fusion procedure on density *(top)*, accuracy and precision *(bottom)*.

fusion method for the generation of 3D city models. The point cloud was meshed in tiles, due to limited memory and faster runtime. However, border discontinuities are introduced by meshing single tiles instead of the entire fused point cloud at once.

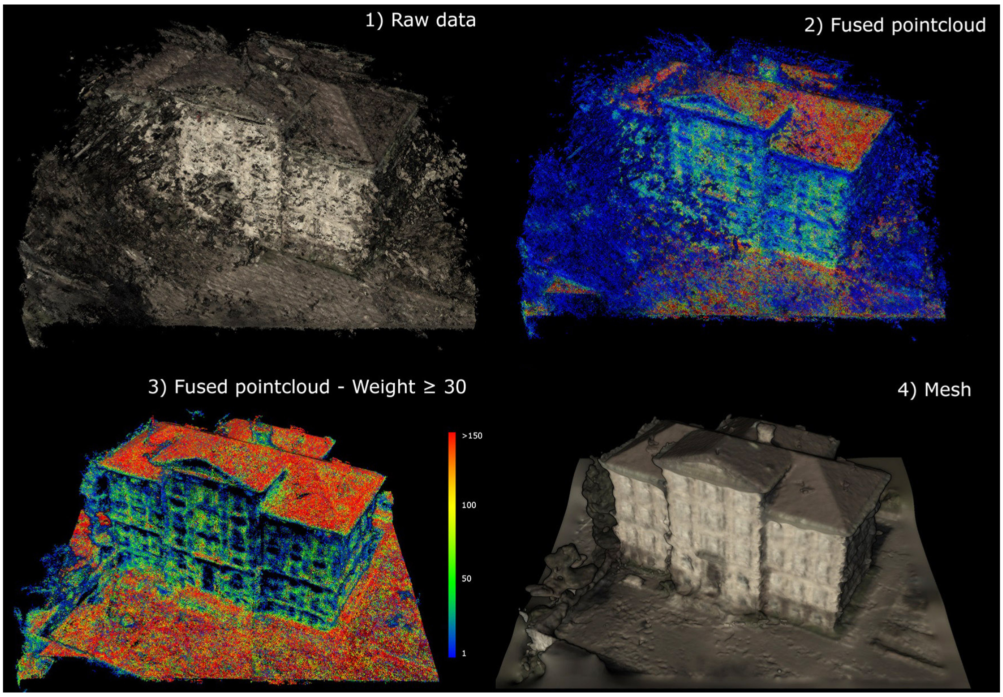


Figure 5.9: Intermediate results of the point cloud fusion method: *(1)* Raw data from dense image matching (50.64 M points), *(2)* fused point cloud (1.73 M points), *(3)* discarded points with weights smaller than $\alpha = 30$ (0.47 M points), *(4)* generated mesh.

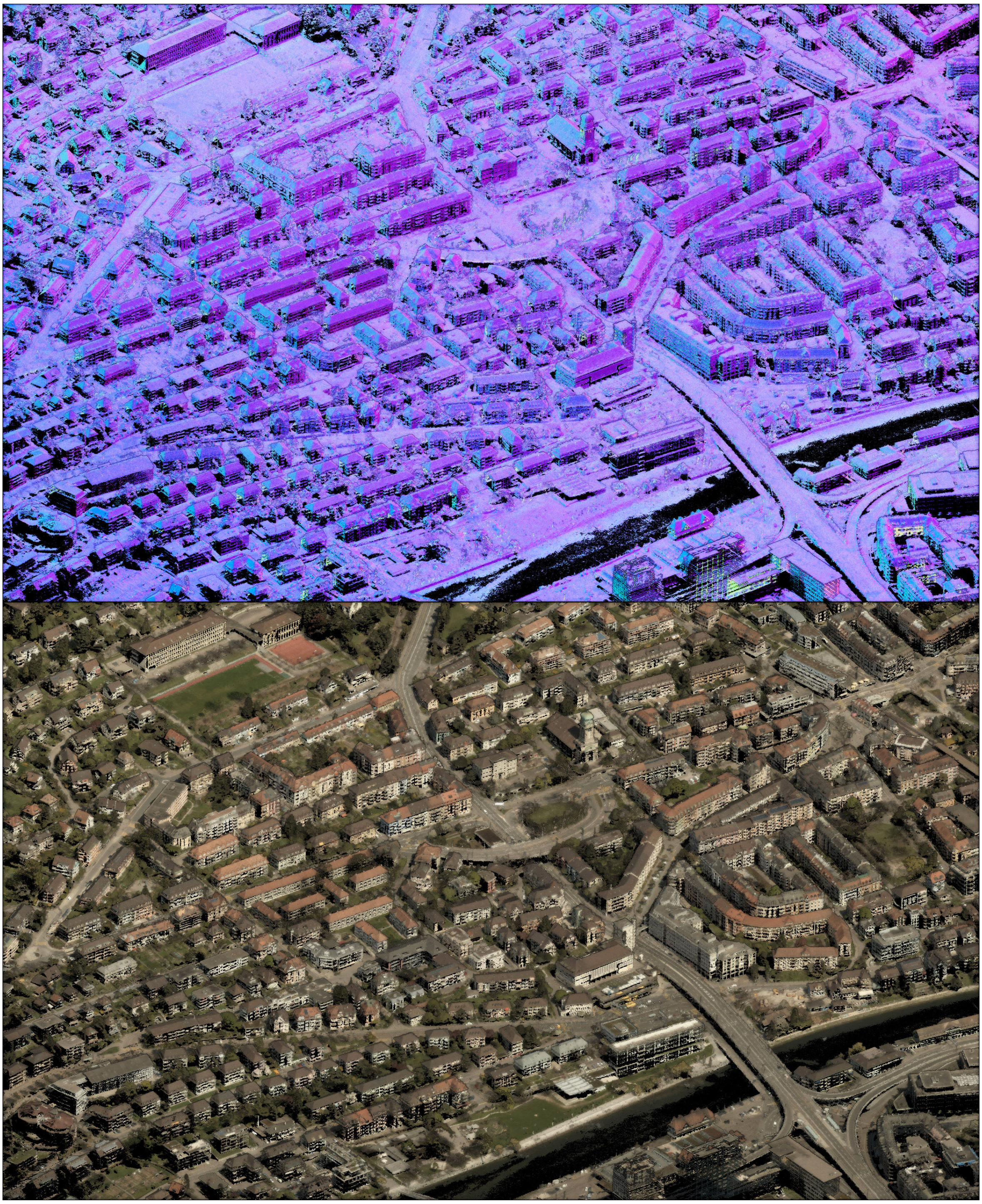


Figure 5.10: Fused point cloud with color-coded normals *(top)* and RGB information *(bottom)*.

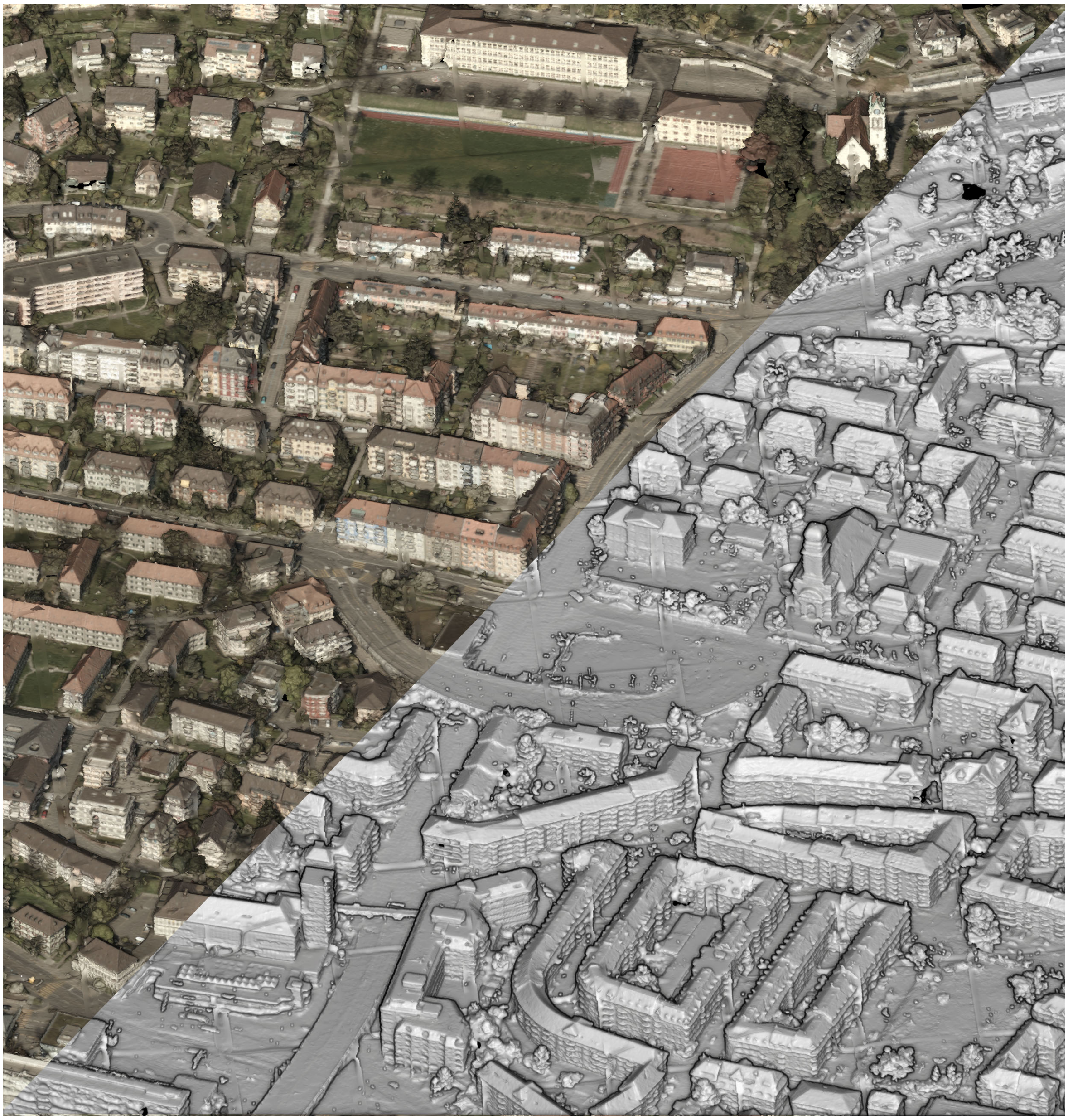


Figure 5.11: Reconstructed mesh from several fused point cloud tiles, showing the capability of the point cloud fusion method for the generation of 3D city models.

5.1.4 Evaluation

The evaluation of the fusion procedure is performed to measure the potential of the proposed method. Figure 5.12 illustrates the capability of the fusion routine to improve the quality of the reconstructed surface visually.

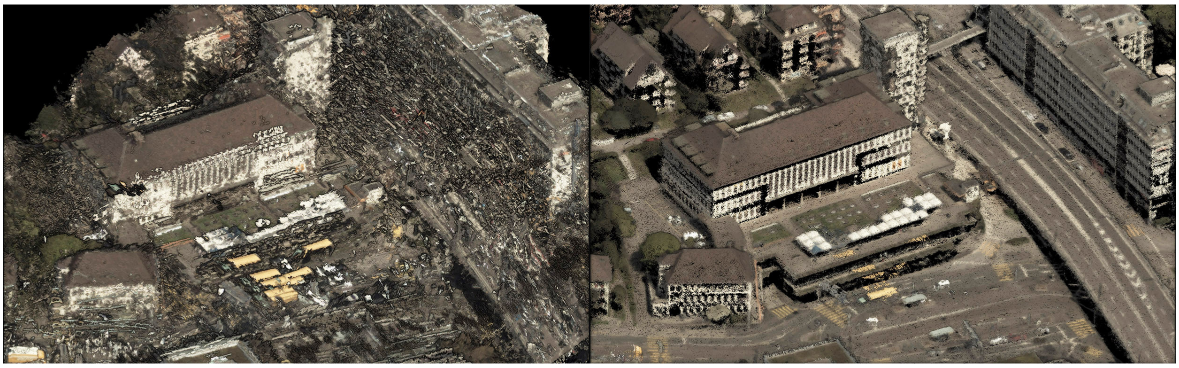


Figure 5.12: Reconstructed point cloud before *(left)* and after *(right)* the fusion routine.

The evaluation is carried out by computing vertical DSMs of different facade patches distributed over the reconstructed scene. In order to measure the potential of the fusion routine different statistical measures were analyzed. The root mean square error (RMSE) of the deviations between the reference point cloud and fused point cloud gives information about the accuracy of the 3D geometry. The standard deviation of the digital surface model indicates the noise level of the point cloud, more specifically the distribution of points perpendicular to the facade. As mentioned before, the density can be controlled by setting the voxel resolution and by regulating the threshold for the minimum weight α. In Table 5.1 the raw point cloud is compared to the fused point cloud considering the influence of TV weights. The minimum weight threshold α is set to generate point clouds with comparable densities. Test areas include the school building located in the northern part of the mapped scene and the tower building located in the south.

Table 5.1: Comparison of the fusion routine regarding weights.

	min. weight α	Density [pnts/m^2]	RMSE Fused PC-TLS [m]	Mean Fused PC-TLS [m]	Std. Dev. of DSM [m]
Raw (unfused)	-	4398.00	0.20	0.11	0.30
Fused (no weights)	20	75.15	0.12	0.07	0.05
Fused (weighted)	30	74.25	0.11	0.06	0.04
Fused (weighted pre-filter TV >1)	18	75.23	0.10	0.05	0.03

Regarding the oblique data set, best results regarding accuracy and performace can be achieved by neglecting points with TV class 1. By doing so, execution time is speed up by a factor of 2.2 (i.e. processing 2.5 billion points on a computer with

8 cores and hyper-threading within a single day and the entire data set within 4 days). Compared to the raw point cloud the fusion procedure reduces noise while improving the accuracy of the point cloud (see Figure 5.13). Considering the registration error of the reference data of 2.2 cm the selected test cases show nearly the same improvement in surface quality. A visual assessment shows that the fused point cloud including all TV classes and applying weights produces the best results regarding completeness and outliers (see Figure 5.14).

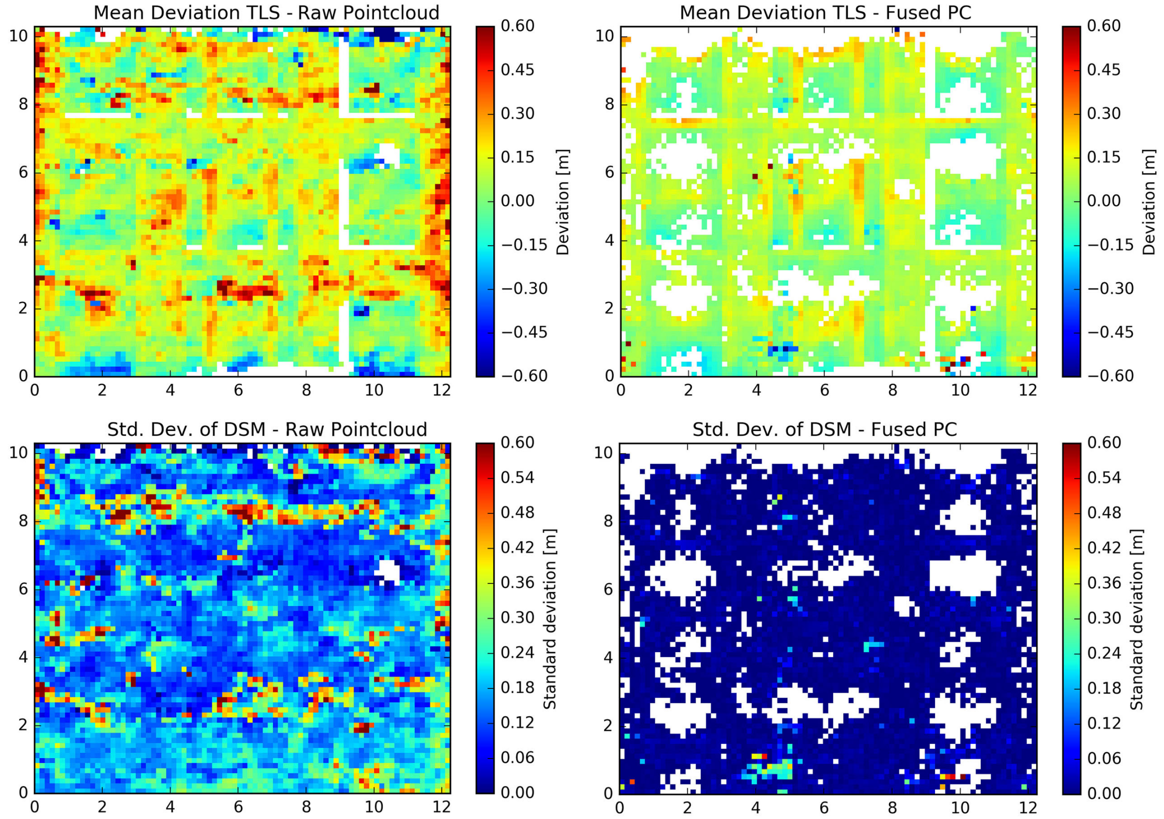


Figure 5.13: Comparison of the main school facade before and after fusion procedure (cf. Figure 5.9): Mean deviation between DSM derived from terrestrial laser scanner data and point cloud *(top)*, and standard deviation of the point clouds DSM representing the level of noise *(bottom)*.

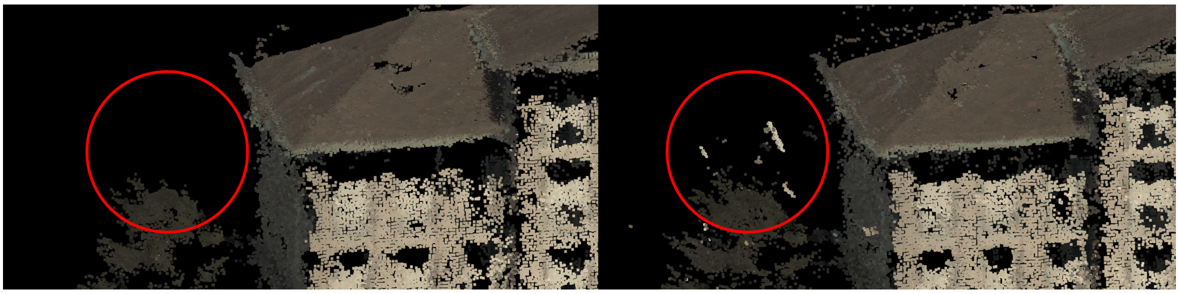


Figure 5.14: Pre-filtering of TV classes: Taking all TV classes into account produces point clouds containing less outliers *(left)*, in contrast to point clouds restricted to TV classes > 1 *(right)*.

As expected, roof structures and other nadir oriented faces are reconstructed with the highest precision and redundancy. Table 5.2 shows that in all cases the precision of the point cloud could be improved while decreasing redundant information. In all cases the noise of the point cloud (i.e. the standard deviation of the DSM) was reduced. Especially for facades with standard deviations of up to 0.5 m, the surface quality could be improved significantly. The mean deviation of the point clouds represents the mean offset (perpendicular to the surface patch) of the point cloud compared to the reference data. As shown in Table 5.2, the mean deviation as well as the RMSE was improved by employing the developed fusion routine. It is worth mentioning that remaining deviations to the reference data (e.g. a global shift of the reconstructed scene) could be induced by image registration errors, which can not be eliminated by the proposed fusion method.

Table 5.2: Comparison of test areas before and after the point cloud fusion.

	Density [pnts/m^2]	RMSE Fused PC-TLS [m]	Mean Fused PC-TLS [m]	Std. Dev. of DSM [m]
School (raw)	4398.0	0.20	0.11	0.30
School (fused)	74.3	0.11	0.06	0.04
Tower South (raw)	2345.9	0.38	0.05	0.54
Tower South (fused)	49.4	0.20	0.00	0.09
Tower North (raw)	1781.4	0.43	-0.22	0.45
Tower North (fused)	45.3	0.20	-0.05	0.07
Tower West (raw)	3570.8	0.35	0.24	0.50
Tower West (fused)	62.7	0.26	0.15	0.16
Roof (raw)	13864.2	0.15	-0.02	0.22
Roof (fused)	178.7	0.12	0.03	0.10

5.2 Nadir Aerial Imagery

The nadir image data set consists of 15 panchromatic images captured over the city of Munich. For the analysis of the fusion routine 9 nadir images were selected, covering an area of approximately 2 km × 2 km (see Figure 5.15). The data set was acquired by a DMC II 230 MP aerial image camera with a GSD of 10 cm. To reconstruct complex 3D structures and densely build-up urban areas of the cityscape, the imagery was acquired with an overlap of 80% along as well as across flight direction.

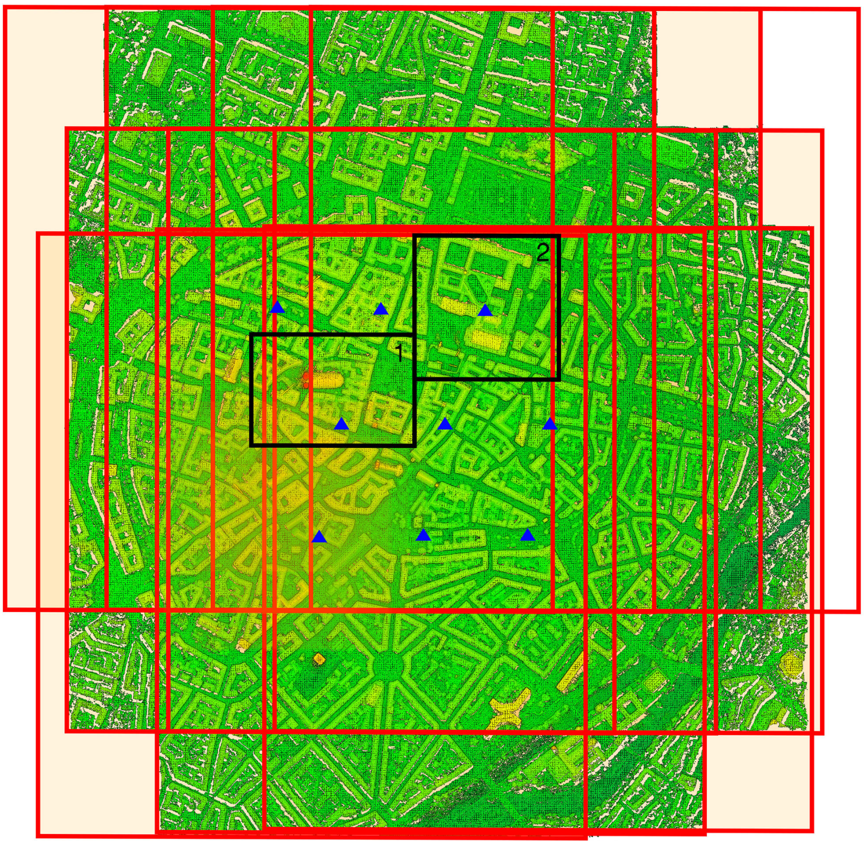


Figure 5.15: Nadir imagery: Camera positions, image footprints and test areas overlaid to the reconstructed scene. Church test area *(1)* and residence test area *(2)*.

Similar to the processing of the oblique data set, the nadir imagery was first oriented with the aid of the provided orientation parameters. Subsequently, the photogrammetric processing in terms of image rectification, stereo matching and spatial point intersection was carried out, resulting in 42 ENH raster files.

In the next step, the disparity maps were assessed using a regularization term of $\tau = 2$, yielding one local total variation based classification per disparity map. Due to the lack of ground truth data, the weight function, which assigns a weight to every TV class, could not be estimated. Hence, the weight function derived from the oblique data set was applied. For the computation of the normal maps a neighborhood size of 5 pixels was defined. After the pre-processing, the dense point clouds and additional information was combined and sorted into 100 m × 100 m tiles.

5.2.1 Point Cloud Fusion

The point cloud fusion routine was applied to the individual tiles using three iterations. The cylinder was defined with a radius of 20 cm (i.e. two times the GSD) and a total height of 3.0 m. The values of the parameters were set by analyzing the raw point cloud (cf. Figure 5.18). The resolution of the octrees leaf nodes, which corresponds to the approximate output resolution, was set to 15 cm. Points with a weight lower than 5 were rejected to eliminate outliers. The minimum weight value was defined by analyzing its impact on the density and presence of remaining outliers, similar to Figure 5.8. The surface reconstruction in terms of mesh generation was performed for fused tiles independently. Hence, discontinuities between tile borders may occur.

Figure 5.18 illustrates the capability of the fusion routine by visualizing the improvement of the surface quality based on the distance between the raw and fused point cloud. The density of the point cloud is reduced significantly by eliminating multiple point redundancy. Moreover, outliers are eliminated by rejecting points with low support (i.e. points with a small accumulated weight value, due to bad coverage and/or high disparity uncertainty). This characteristic is also depicted in Figure 5.18, which shows the raw and fused point cloud as well as the reconstructed mesh.

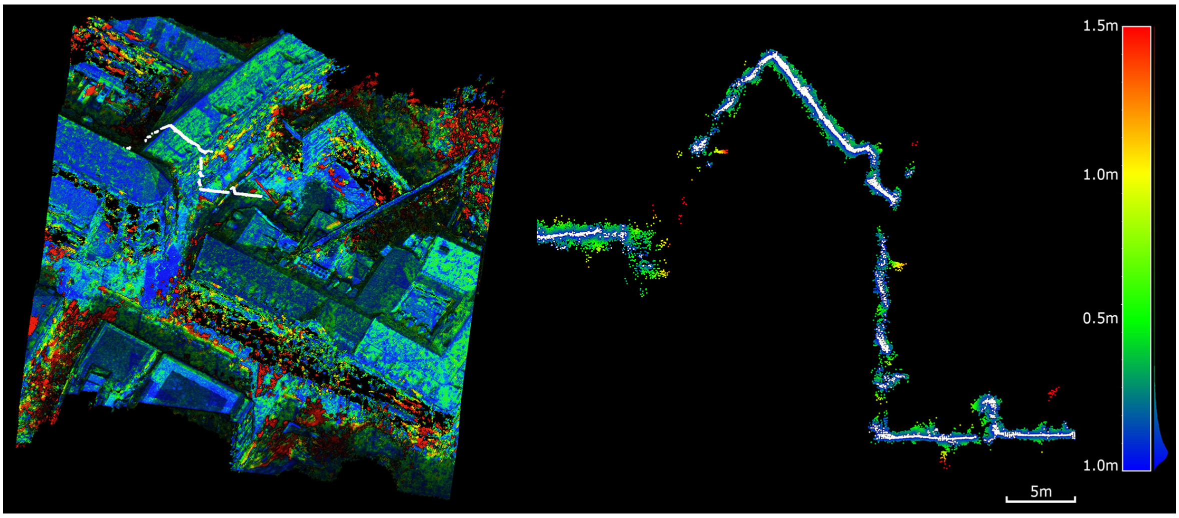


Figure 5.16: Analysis of fused *(white)* and raw *(color-coded)* point cloud based on cloud-to-cloud distance. Overview *(left)* and cross section with a width of 30 cm *(right)*.

Due to the wide angle of the camera, sufficient information is captured to reconstruct facade information. However, facades and other vertical surfaces are distorted by outliers and noise, caused by slanted views due to the nadir configuration and fronto-parallel errors originating from roof points protruding into facades beneath. The fusion routine is able to reduce this point cloud artifacts. Although the fusion method partially eliminates accurate points on facades and other vertical structures, enough information is preserved to generate 3D city models from nadir aerial imagery. As depicted in Figure 5.15, the church test area is located in the western part of the data set. Therefore, no facade information on west facing surfaces can be reconstructed (cf. Figure 5.16, Figure 5.18). The same applies to the residence test area, which is located in the northern part of the data set (see Figure 5.17). Assuming a larger data set, covering the scene from all sides (due to the wide angle of the camera lens) and with high redundancy, the generation of complete 3D city models from nadir imagery is possible.

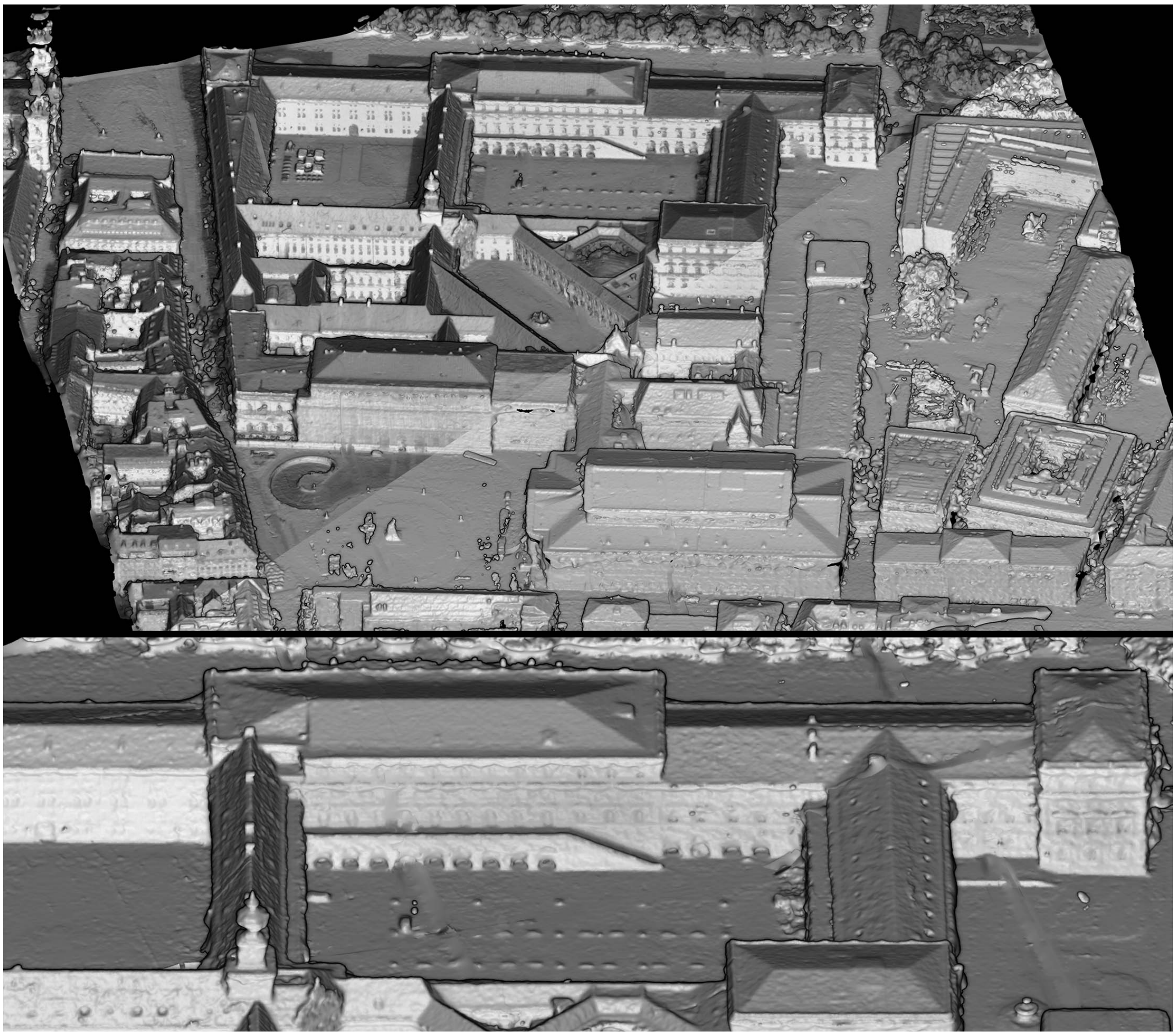


Figure 5.17: Reconstructed point cloud of the residence test area: Mesh and panchromatic colorized mesh *(top)*, close-up view on south facing facades *(bottom)*.

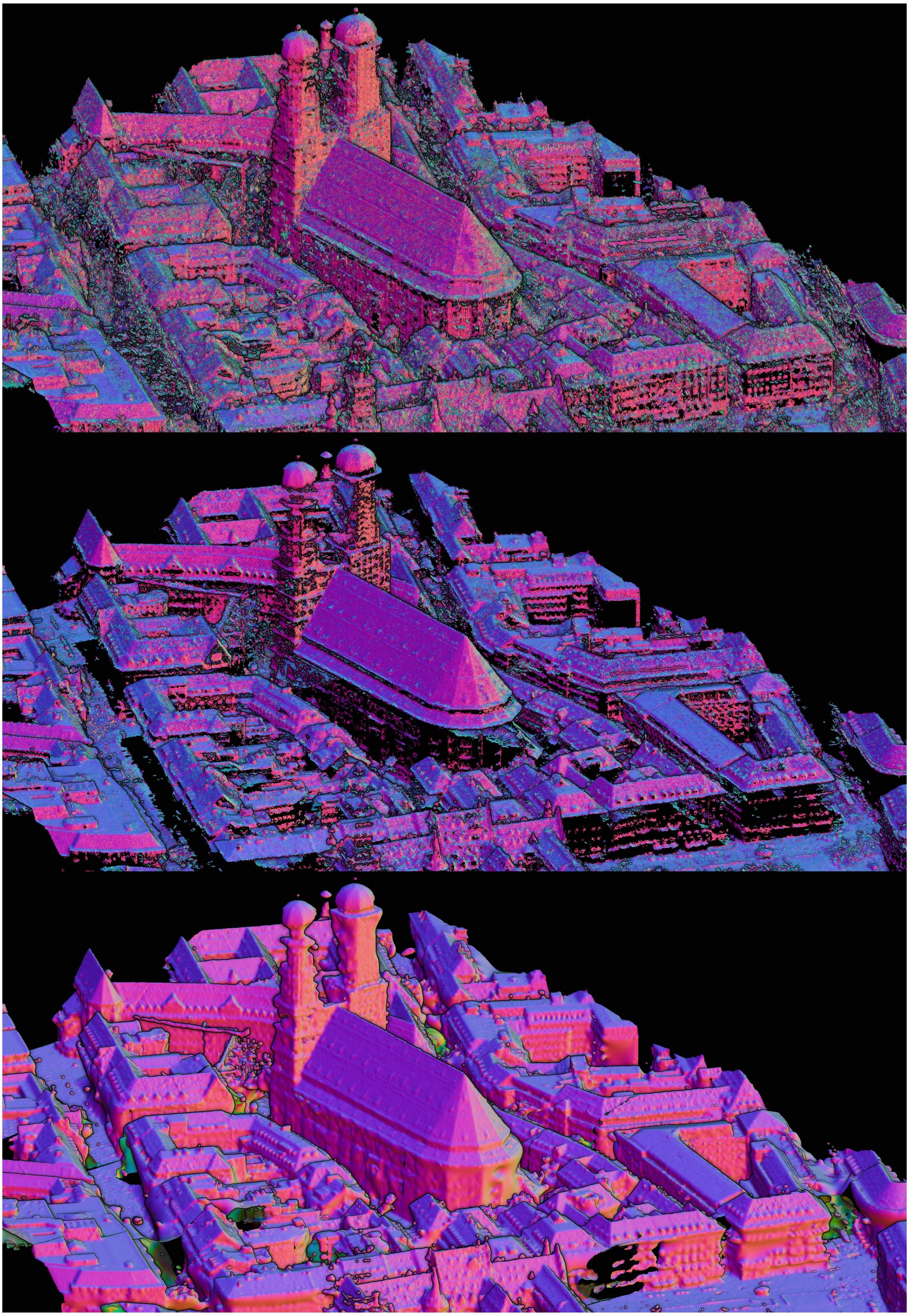


Figure 5.18: Reconstructed point cloud of the church test area: Raw dense point cloud (317 million points) *(top)*, fused and filtered point cloud (18 million points) *(middle)* and reconstructed mesh using Poisson surface reconstruction *(bottom)*.

5.3 Satellite Imagery

The satellite imagery data set considered for the last evaluation of the fusion procedure was captured by Pléiades satellites. The Pléiades satellite system is a dual system which comprises two identical satellites Pléiades-1A and Pléiades-1B providing very high resolution (VHR) imagery. The sensors are able to acquire optical panchromatic (0.7 m GSD) and multispectral (2.8 m GSD) imagery. Their ability to change the pointing angle in a range of ± 47° (standard mode ± 30°) enables the generation of high-resolution DSMs.

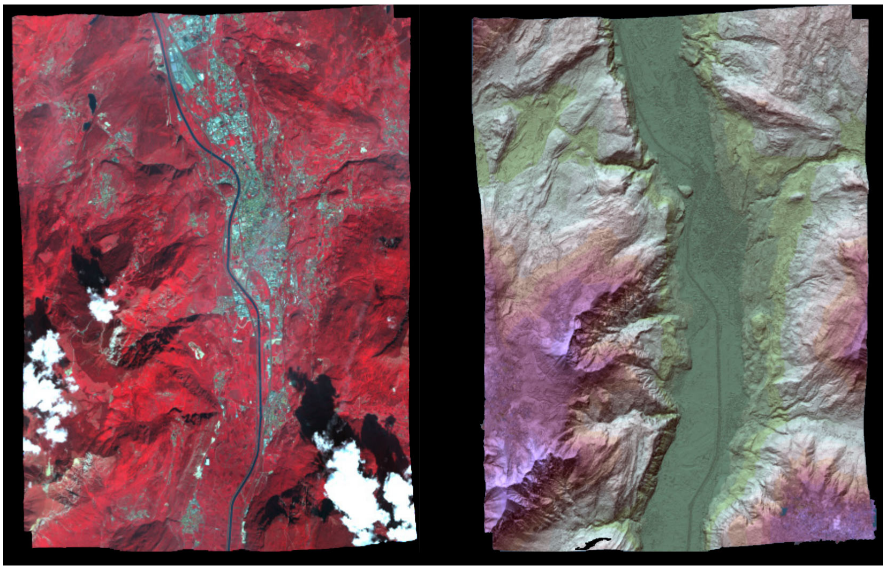


Figure 5.19: Trento triplet data: Orthophoto *(left)* relief shaded DSM *(right)* by Perko et al., 2014, p. 2.

The test data set consists of three scenes and was acquired over the region of Trento (Italy) covering rural as well as mountainous terrain of approximate 220 km^2 (see Figure 5.19). As mentioned by Perko et al. (2014), the Pléiades-1A triplet data is not optimal for 3D reconstruction, since the first stereo pair has a very small intersection angle of 5.5°, while the intersection angle of the second stereo pair is huge (27.3°).

The following evaluation builds upon the work of Perko et al. (2014) who assessed the mapping potential of Pléiades stereo and triplet data. In their work, the mentioned test site located in Trento was processed with the aid of RSG yielding 6 DSMs which were then fused. For this work, the ENH raster files generated in the course of 3D reconstruction by Perko et al. (2014), are applied to the proposed fusion routine. For more details on the test data, sensor model optimization and 3D reconstruction (i.e. Epipolar rectification, image matching and spatial point intersection) of the Pléiades triplet data, the reader is referred to Perko et al. (2014).

5.3.1 Point Cloud Fusion

Prior to the point cloud fusion, the pre-processing steps were applied and the point cloud was sorted into 500 m × 500 m tiles holding a total of 4 billion points. Due to the dynamic image geometry of the optical line scanner, the camera position, which is required to ensure correct normal orientation, is approximated by the mean orbit position during image acquisition.
The actual fusion was carried out in three iterations using a cylinder radius of 0.75 m and a total height of 5 m. The height of the cylinder was defined by analyzing the noise of the raw point cloud. The approximate output density was set to 0.5 m which corresponds to the GSD of the input imagery. It is worth mentioning, that the original GSD of the imagery differs between 0.7 and 0.77 m, but is delivered with an upsampled GSD of 0.5 m. Similar to Section 5.2, the weighting function calculated from the oblique imagery data set and regularization term of $\tau = 2$, was used for the satellite imagery. However, the underlying sensor geometry and the imagery itself differs between both data sets. Hence, the results obtained from the fusion routine could be further improved by assessing the correlation between TV classes and disparity uncertainty particular for Pléiades imagery data sets.

The fusion of the entire data set consisting of 4 billion points, with an approximate output GSD of 0.5 m, requires 72 hours on a computer with 8 cores and hyperthreading. Whereas the generation of an overview point cloud, by fusing the entire data set with an approximate output GSD of 2 m, can be computed in 4 hours. It is worth mentioning that increasing the redundancy by adding additional imagery, does not significantly affect the overall runtime of the fusion procedure, due to the concept of initial pointset creation.

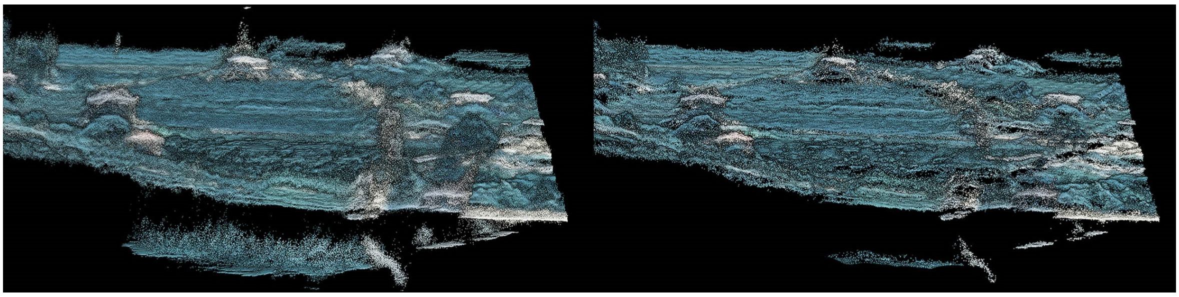


Figure 5.20: Comparison between raw *(left)* and fused *(right)* point cloud. Due to low redundancy, fronto-parallel errors, can not be eliminated completely.

Figure 5.20 shows the point clouds of a selected tile before and after applying the fusion method. While a decrease in the noise of the point cloud can be verified, outliers, especially fronto-parallel errors, aggravated by the unfavorable stereo configuration (i.e. bad angle of intersection), can not be eliminated completely. To mitigate this issue, additional image scenes may be used to increase the support (i.e. in terms of a higher accumulated weight) on correctly reconstructed points.

In order to further analyze the results obtained from the fusion procedure a DSM was computed for a selected test area and compared to the results presented by Perko et al. (2014). The test site covers a hospital building and surrounding area of 360 m × 360 m. The DSM derived from the fused point cloud is generated by calculating the mean height of points located in the same raster cell. The DSM is compared to the results presented in Perko et al. (2014) and a DSM derived from LiDAR data. Due to the large temporal gap between the LiDAR and Pléiades data acqusition and infrastructural changes like construction activities, no quantitative analysis was carried out for the selected test area. Figure 5.21 shows the different DSMs with terrain heights scaled between 240 m (black) and 300 m (white).

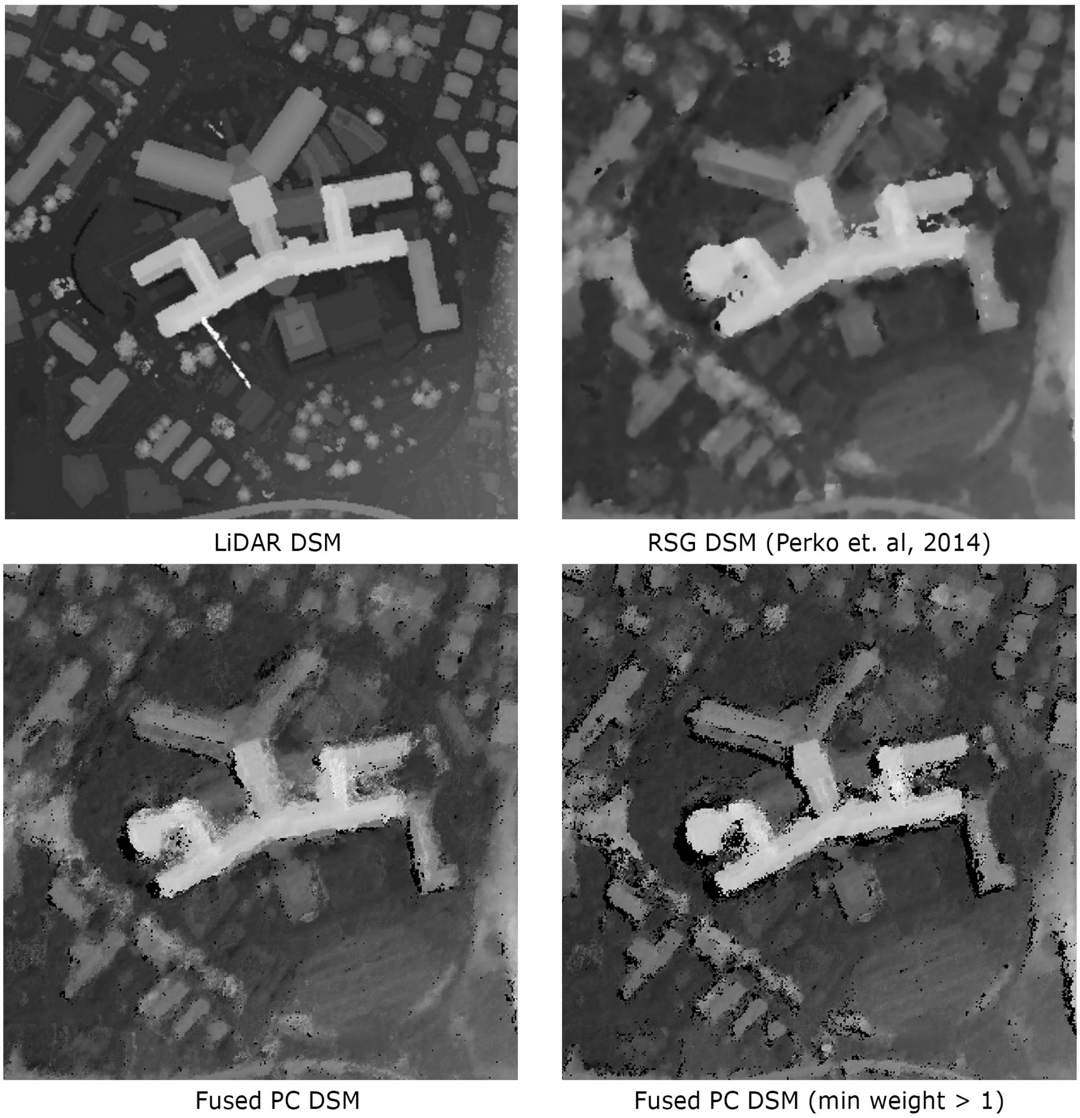


Figure 5.21: Detailed view of the hospital test area: DSM derived from the fused point cloud compared to the DSM presented by Perko et al. (2014) and a LiDAR DSM. Terrain heights are scaled from 240 *(black)* to 300 *(white)* meters.

While the DSM derived from the entire fused point cloud shows a high variance in the terrain heights (see Figure 5.21 *bottom left*), the point cloud restricted to a minimum weight greater than 1 (*bottom right*) represents a smoother surface representation. Especially in the DSM with discarded weights (*bottom right*), data gaps are present in areas with bad coverage or occlusions caused by the stereo configuration. Compared to the DSM generated with the aid of RSG, the fused point cloud DSM presevers 3D breaklines better and the shape of the reconstructed hospital building looks more similar w.r.t. the LiDAR model. This is possible due to the elimination of outliers, especially in areas with rapidly changing heights. Whereas the RSG DSM shows a flattening behavior of the height values in these discontinuous areas. However, for a detailed analysis of the DSMs and its precision a quantitative evaluation would be necessary.

Surface reconstruction in terms of meshed surface generation was carried out for the hospital test area and a larger subset of the data set (see Figure 5.22 and 5.23). The depicted scenes shows the potential of Pléiades satellite imagery for the generation of 3D city models based on the proposed fusion strategy. Although facade information and small buildings are not reconstructed completely, sufficient information is obtained to generate city models with a Level of detail (LOD) of 1 up to 2, according to the CityGML[2] LOD scheme. The Poisson surface reconstruction algorithm, utilized for the generation of the mesh representation, usually maintains sharp features in areas of dense sampling and provides smooth fitting in sparsely sampled regions (Kazhdan and Hoppe, 2013). This characteristic is observable in Figure 5.23 and 5.22 on buildings where no facade information is reconstructed. Therefore, 3D city models with higher quality may be generated by applying different surface reconstruction methods on the fused and oriented point cloud. Finally, the visual quality of the meshed surface representation may be increased by employing texture mapping algorithms to colorize the 3D model, rather than simply colorizing the faces of the polygon mesh.

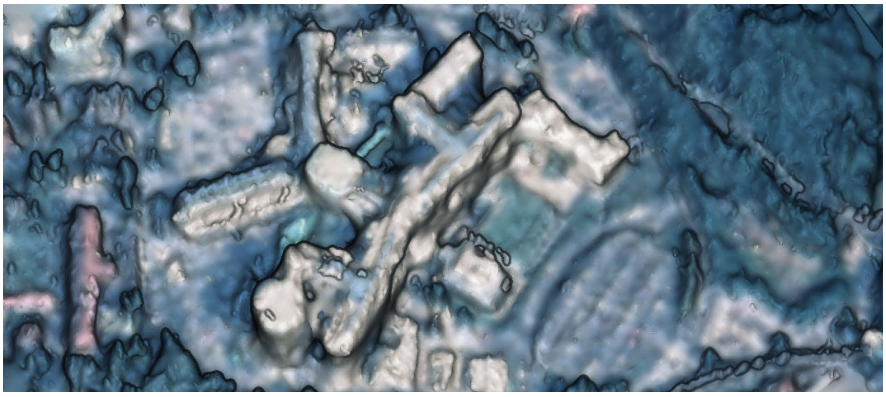


Figure 5.22: Reconstructed mesh of the hospital test area.

[2] `https://www.citygml.org/`

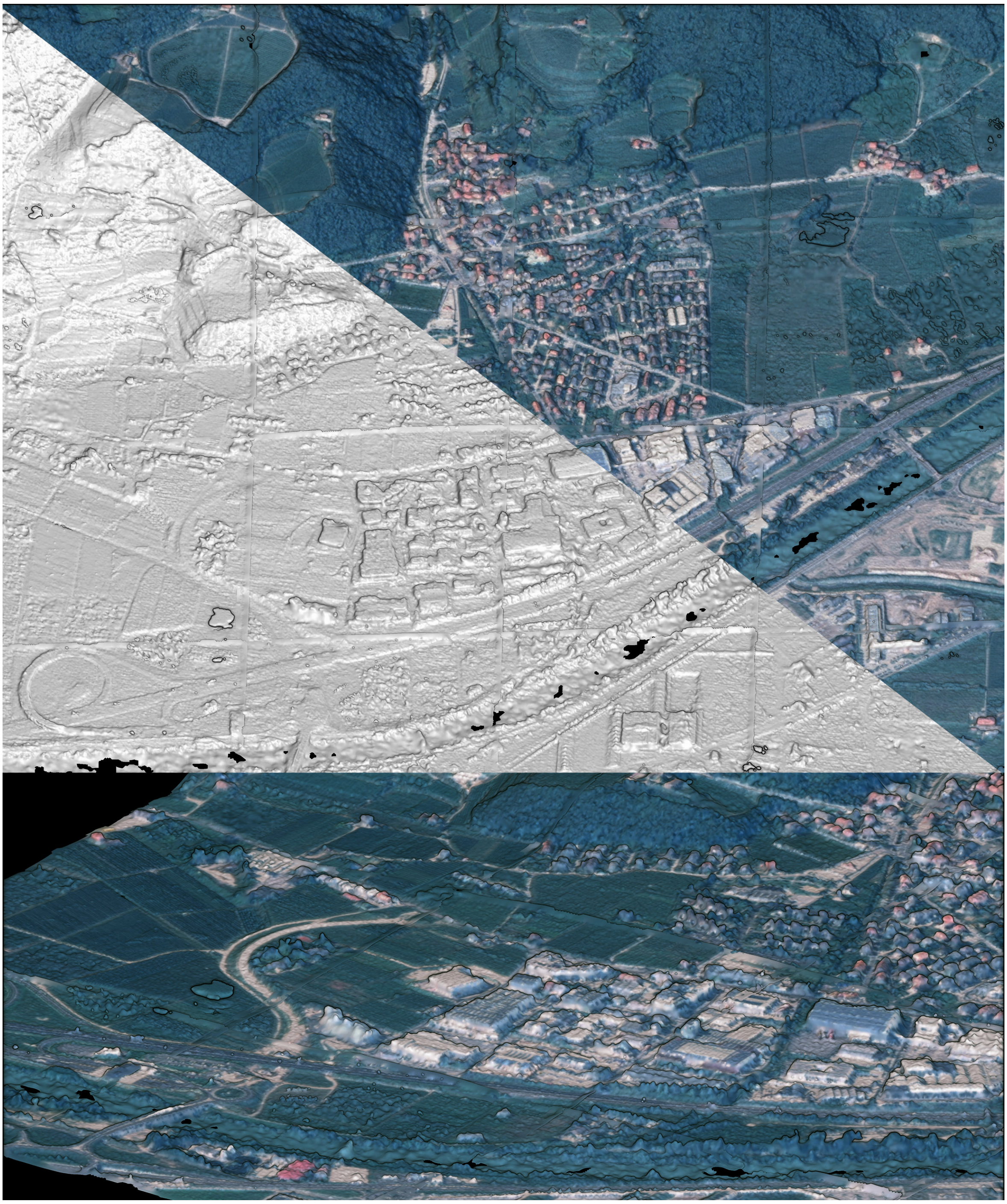


Figure 5.23: Reconstructed mesh from several fused point cloud tiles. Colorized mesh faces based on multispectral information of the input imagery.

Figure 5.24 as well as Figure 5.25 depict an overview of the reconstructed scene.

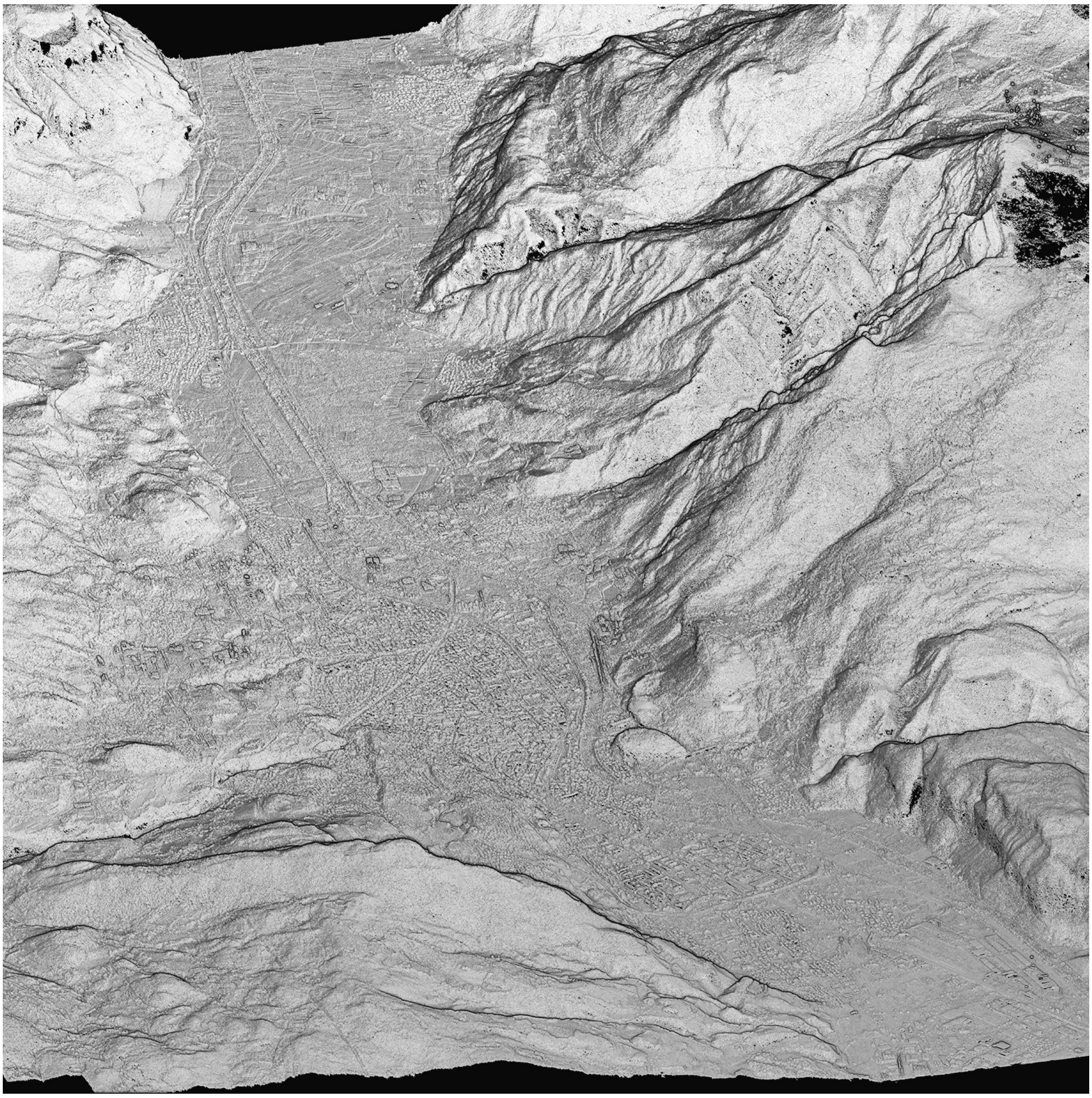


Figure 5.24: Overview of the reconstructed scene (view along the valley). Partially undetected outliers and missing points due to clouds in the input imagery *(top right)* (cf. Figure 5.19)

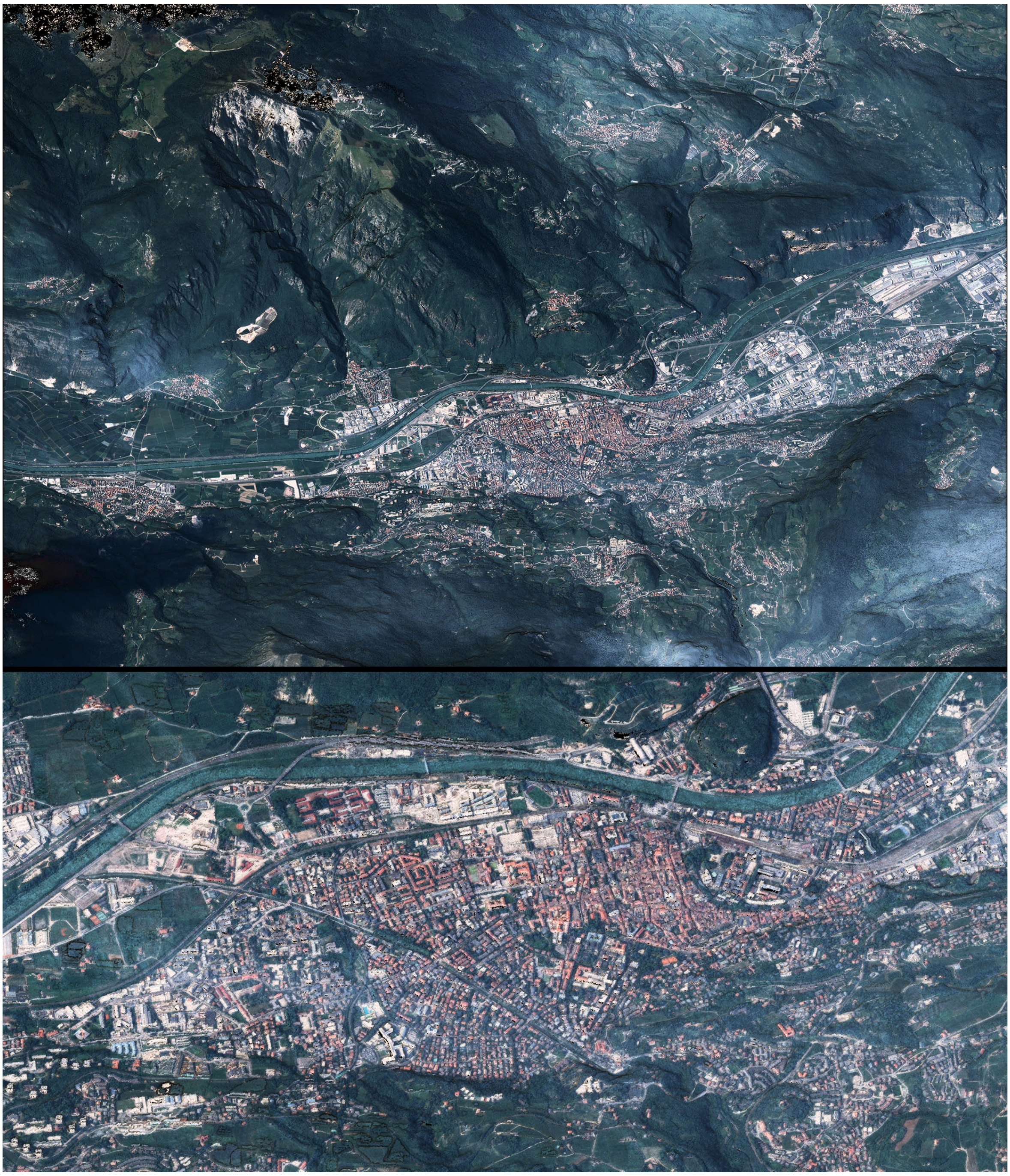


Figure 5.25: Overview of the reconstructed scene. Underlying colorized point cloud fused for visualization purpose within 4 hours by setting the approximate output GSD to 2 m.

6 Summary and Outlook

The object of this thesis was the development of a scalable point cloud fusion strategy which fuses multiple stereo derived point clouds into one consistent surface representation. In the first part, a basic overview of 3D reconstruction methods from multiple images, with focus on depth map based multi-view stereo (MVS) system, has been given. Moreover, recent work considering the generation of consistent 3D surface representations from imagery, especially large-scale data sets, was discussed.

The next part of the thesis focused on the proposed fusion strategy, comprising details of the implemented pipeline. The fusion strategy was developed based on the results of the literature review by following a Divide and Conquer approach to ensure scalability. The fusion strategy itself builds on the concept of consecutive fusion of points along defined filtering directions and utilizes a total variation based classification of the disparity maps to improve the quality of the reconstructed scene. Once the underlying 3D reconstruction system was described, pre-processing steps, required input parameters and the fusion strategy were explained. The potential of the fusion method was evaluated based on oblique imagery. For further analysis of the proposed method, the applicability of the developed algorithm was tested on imagery with different spatial resolutions and stereo configuration (nadir airborne and satellite imagery). In case of the oblique data set, detailed numerical analysis have been conducted with the aid of terrestrial laser scanner reference data. The results demonstrated that the fusion strategy is applicable for the generation of consistent surface representations of large data sets consisting of billions of points. Results showed that, in any case, redundancy of the point cloud could be decreased while improving the overall surface quality in terms of eliminated outliers and reduced point cloud noise. Moreover, the fused and oriented point clouds are suitable for meshing algorithms generating 3D models.

The analysis showed that the fusion procedure performs best with data sets captured with good coverage and high overlaps. Processing data sets with low redundancy (e.g. satellite imagery) appears to be problematic since not all outliers may be detected due to low support of correctly reconstructed points. The underlying stereo method used in this work is based on spatial point intersection yielding two disparity maps per stereo pair. Therefore, results may be improved by employing multi-baseline triangulation algorithms which lead to one depth map per input image. Thus, outliers and noise may be reduced significantly by eliminating errors in the process of 3D reconstruction prior to the fusion. Currently, the fusion method is based on a static octree data structure producing point clouds with nearly consistent density. Further improvements may include the implementation of a dynamic octree structure to reconstruct scenes with multiple scales. Furthermore, utilizing an out-of-core octree implementation would eliminate the need for the external point cloud tiling step. Finally, the quality of reconstructed mesh representation could be improved by stitching meshed tiles and employing texture mapping algorithms to colorize the 3D model based on the input imagery.

Abbreviations

DSM	Digital surface model
DTM	Digital terrain model
ENH	East-North-Height
GPS	Global Positioning System
GSD	Ground sampling distance
IMU	Inertial measurement unit
LiDAR	Light detection and ranging
LOD	Level of detail
MP	Megapixel
MVS	Multi-view stereo
PC	Point cloud
RMSE	Root mean square error
RSG	Remote Sensing Package Graz
RTK	Point cloud
SAR	Synthetic aperture radar
SFM	Structure from motion
SGM	Semi-Global Matching
TLS	Terrestrial laser scanner
TV	Total variation
UAV	Unmanned aerial vehicle
VHR	Very high resolution
VRIP	Volumetric range image processing

List of Figures

List of Tables

Bibliography

Ahmadabadian, A. H., S. Robson, J. Boehm, M. Shortis, K. Wenzel, and D. Fritsch (2013). "A comparison of dense matching algorithms for scaled surface reconstruction using stereo camera rigs." In: *ISPRS Journal of Photogrammetry and Remote Sensing* 78, pp. 157–167. ISSN: 0924-2716. DOI: http://dx.doi.org/10.1016/j.isprsjprs.2013.01.015. URL: http://www.sciencedirect.com/science/article/pii/S0924271613000452 (cit. on p. 24).

Akio, D. and A. Koide (1991). "An efficient method of triangulating equi-valued surfaces by using tetrahedral cells." In: *IEICE TRANSACTIONS on Information and Systems* 74.1, pp. 214–224 (cit. on p. 15).

Amenta, N., S. Choi, and R. K. Kolluri (2001). "The Power Crust." In: *Proceedings of the Sixth ACM Symposium on Solid Modeling and Applications*. SMA '01. Ann Arbor, Michigan, USA: ACM, pp. 249–266. ISBN: 1-58113-366-9. DOI: 10.1145/376957.376986. URL: http://doi.acm.org/10.1145/376957.376986 (cit. on p. 14).

Baumgart, B. G. (1974). "Geometric Modeling for Computer Vision." AAI7506806. PhD thesis. Stanford, CA, USA (cit. on p. 13).

Cavegn, S., N. Haala, S. Nebiker, M. Rothermel, and P. Tutzauer (2014). "Benchmarking High Density Image Matching for Oblique Airborne Imagery." In: *ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 45–52. DOI: 10.5194/isprsarchives-XL-3-45-2014 (cit. on pp. 24, 32).

Curless, B. and M. Levoy (1996). "A Volumetric Method for Building Complex Models from Range Images." In: *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '96. New York, NY, USA: ACM, pp. 303–312. ISBN: 0-89791-746-4. DOI: 10.1145/237170.237269. URL: http://doi.acm.org/10.1145/237170.237269 (cit. on p. 15).

Deuber, M. (2014). "Oblique Photogrammetry – Dense Image Matching mit Schrägluftbildern." MA thesis. FHNW Fachhochschule Nordwestschweiz (cit. on p. 32).

Fuhrmann, S. and M. Goesele (2011). "Fusion of Depth Maps with Multiple Scales." In: *ACM Trans. Graph.* 30.6, 148:1–148:8. ISSN: 0730-0301. DOI: 10.1145/2070781.2024182. URL: http://doi.acm.org/10.1145/2070781.2024182 (cit. on pp. 15, 16).

Furukawa, Y. and J. Ponce (2010). "Accurate, Dense, and Robust Multiview Stereopsis." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.8, pp. 1362–1376. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2009.161 (cit. on p. 13).

Furukawa, Y. and C. Hernández (2015). "Multi-View Stereo: A Tutorial." In: *Foundations and Trends in Computer Graphics and Vision* 9.1-2, pp. 1–148. ISSN: 1572-2740. DOI: 10.1561/0600000052. URL: http://dx.doi.org/10.1561/0600000052 (cit. on pp. 8, 13).

Galliani, S., K. Lasinger, and K. Schindler (2015). "Massively Parallel Multiview Stereopsis by Surface Normal Diffusion." In: *IEEE International Conference on Computer Vision (ICCV)* (cit. on pp. 16, 28).

Hartley, R. I. and A. Zisserman (2004). *Multiple View Geometry in Computer Vision*. Second. Cambridge University Press, ISBN: 0521540518 (cit. on p. 11).

Hilton, A. and J. Illingworth (1997). "Multi-resolution geometric fusion." In: *Proceedings. International Conference on Recent Advances in 3-D Digital Imaging and Modeling (Cat. No.97TB100134)*, pp. 181–188. DOI: 10.1109/IM.1997.603864 (cit. on p. 15).

Hirschmüller, H. (2008). "Stereo Processing by Semiglobal Matching and Mutual Information." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.2, pp. 328–341. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2007.1166 (cit. on p. 14).

Hoppe, H., T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle (1992). "Surface Reconstruction from Unorganized Points." In: *Proceedings of the 19th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '92. New York, NY, USA: ACM, pp. 71–78. ISBN: 0-89791-479-1. DOI: 10.1145/133994.134011. URL: http://doi.acm.org/10.1145/133994.134011 (cit. on p. 14).

Isenburg, M. (2013). "LASzip." In: *Photogrammetric Engineering and Remote Sensing* 79.2, pp. 209–217. ISSN: 0099-1112. DOI: doi:10.14358/PERS.79.2.209 (cit. on p. 25).

Kazhdan, M. and H. Hoppe (2013). "Screened Poisson Surface Reconstruction." In: *ACM Trans. Graph.* 32.3, 29:1–29:13. ISSN: 0730-0301. DOI: 10.1145/2487228.2487237. URL: http://doi.acm.org/10.1145/2487228.2487237 (cit. on pp. 16, 28, 35, 49).

Kuhn, A., H. Huang, M. Drauschke, and H. Mayer (2016a). "Fast Probabilistic Fusion of 3D Point Clouds via Occupancy Grids for Scene Classification." In: *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* III-3, pp. 325–332. DOI: 10.5194/isprs-annals-III-3-325-2016. URL: http://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/III-3/325/2016/ (cit. on p. 16).

Kuhn, A. (2014). "Scalable 3D Surface Reconstruction by Local Stochastic Fusion of Disparity Maps." PhD thesis. Neubiberg: Universität der Bundeswehr München, Fakultät für Informatik (cit. on pp. 1, 5, 11, 14–16, 20–23).

Kuhn, A., H. Hirschmüller, and H. Mayer (2013). "Multi-Resolution Range Data Fusion for Multi-View Stereo Reconstruction." In: *Pattern Recognition: 35th German Conference, GCPR 2013, Saarbrücken, Germany, September 3-6, 2013. Proceedings*. Ed. by J. Weickert, M. Hein, and B. Schiele. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 41–50. ISBN: 978-3-642-40602-7. DOI: 10.1007/978-3-642-40602-7_5. URL: http://dx.doi.org/10.1007/978-3-642-40602-7_5 (cit. on pp. 15, 16).

Kuhn, A., H. Hirschmüller, D. Scharstein, and H. Mayer (2016b). "A TV Prior for High-Quality Scalable Multi-View Stereo Reconstruction." In: *International Journal of Computer Vision*, pp. 1–16. ISSN: 1573-1405. DOI: 10.1007/s11263-016-0946-x. URL: http://dx.doi.org/10.1007/s11263-016-0946-x (cit. on pp. 16, 21).

Kutulakos, K. N. and S. M. Seitz (1999). "A theory of shape by space carving." In: *Proceedings of the Seventh IEEE International Conference on Computer Vision*. Vol. 1, 307–314 vol.1. DOI: 10.1109/ICCV.1999.791235 (cit. on p. 13).

Labatut, P., J.-P. Pons, and R. Keriven (2009). "Robust and Efficient Surface Reconstruction From Range Data." In: *Computer Graphics Forum*. ISSN: 1467-8659. DOI: 10.1111/j.1467-8659.2009.01530.x (cit. on p. 15).

Lorensen, W. E. and H. E. Cline (1987). "Marching Cubes: A High Resolution 3D Surface Construction Algorithm." In: *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '87. New York, NY, USA: ACM, pp. 163–169. ISBN: 0-89791-227-6. DOI: 10.1145/37401.37422. URL: http://doi.acm.org/10.1145/37401.37422 (cit. on p. 15).

Luhmann, T. (2010). *Nahbereichsphotogrammetrie Grundlagen, Methoden und Anwendungen*. 3rd ed. Wichmann Verlag, Heidelberg. ISBN: 978-3-87907-479-2 (cit. on pp. 9–11).

McGlone, J. C. (2013). *Manual of photogrammetry*. English. 6th ed. American Society for Photogrammetry and Remote Sensing. ISBN: 1-57083-099-1 (cit. on pp. 8, 11).

Molton, N. and M. Brady (2000). "Practical Structure and Motion from Stereo When Motion is Unconstrained." In: *International Journal of Computer Vision* 39.1, pp. 5–23. ISSN: 1573-1405. DOI: 10.1023/A:1008191416557. URL: http://dx.doi.org/10.1023/A:1008191416557 (cit. on p. 20).

Pajarola, R. (1998). "Large scale terrain visualization using the restricted quadtree triangulation." In: *Visualization*, pp. 19–26. DOI: 10.1109/VISUAL.1998.745280 (cit. on p. 19).

Perko, R., H. Raggam, K. Gutjahr, and M. Schardt (2014). "Assessment of the mapping potential of Pleiades stereo and triplet data." In: *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* II-3, pp. 103–109 (cit. on pp. 12, 46, 48).

Pierrot-Deseilligny, M. and N. Paparoditis (2006). "A multiresolution and optimization-based image matching approach: An application to surface reconstruction from spot5-hrs stereo imagery." In: *In: Proc. of the ISPRS Conference Topographic Mapping From Space (With Special Emphasis on Small Satellites), ISPRS* (cit. on p. 14).

Piracés, P. G. I. (2008). "Contributions to the Bayesian Approach to Multi-View Stereo." Theses. Institut National Polytechnique de Grenoble - INPG. URL: https://tel.archives-ouvertes.fr/tel-00279825 (cit. on p. 12).

Rothermel, M., N. Haala, and D. Fritsch (2016). "A Median-Based Depthmap Fusion Strategy for the Generation of Oriented Points." In: *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 115–122. DOI: 10.5194/isprs-annals-III-3-115-2016 (cit. on pp. 7, 16, 19, 25–28).

Rothermel, M. (2017). "Development of a SGM-based multi-view reconstruction framework for aerial imagery." PhD thesis. University of Stuttgart, Germany. URL: http://nbn-resolving.de/urn:nbn:de:bsz:93-opus-ds-90675 (cit. on p. 6).

Rusu, R. B. and S. Cousins (2011). "3D is here: Point Cloud Library (PCL)." In: *IEEE International Conference on Robotics and Automation (ICRA)*. Shanghai, China (cit. on p. 26).

Schindler, K. (2014). "Mathematical Foundations of Photogrammetry." In: *Handbook of Geomathematics*. Ed. by W. Freeden, M. Z. Nashed, and T. Sonar. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–14. ISBN: 978-3-642-27793-1. DOI: 10.1007/978-3-642-27793-1_63-1. URL: http://dx.doi.org/10.1007/978-3-642-27793-1_63-1 (cit. on pp. 5, 10).

Schönfelder, A., R. Perko, K. Gutjahr, and M. Schardt (2017). "Fusion of Point Clouds derived from Aerial Images." In: *OAGM and ARW Joint Workshop*. 2. Wien, Austria, pp. 139–144. DOI: 10.3217/978-3-85125-524-9 (cit. on p. 3).

Seitz, S. M., B. Curless, J. Diebel, D. Scharstein, and R. Szeliski (2006). "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 1, pp. 519–528. DOI: 10.1109/CVPR.2006.19 (cit. on pp. 12–15).

Seitz, S. M. and C. R. Dyer (1999). "Photorealistic Scene Reconstruction by Voxel Coloring." In: *Int. J. Comput. Vision* 35.2, pp. 151–173. ISSN: 0920-5691. DOI: 10.1023/A:1008176507526. URL: http://dx.doi.org/10.1023/A:1008176507526 (cit. on p. 13).

Turk, G. and M. Levoy (1994). "Zippered Polygon Meshes from Range Images." In: *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '94. New York, NY, USA: ACM, pp. 311–318. ISBN: 0-89791-667-0. DOI: 10.1145/192161.192241. URL: http://doi.acm.org/10.1145/192161.192241 (cit. on p. 15).

Ummenhofer, B. and T. Brox (2015). "Global, Dense Multiscale Reconstruction for a Billion Points." In: *IEEE International Conference on Computer Vision (ICCV)*. URL: http://lmb.informatik.uni-freiburg.de//Publications/2015/UB15 (cit. on p. 16).

Vogiatzis, G., P. H. S. Torr, S. M. Seitz, and R. Cipolla (2008). "Reconstructing Relief Surfaces." In: *Image Vision Comput.* 26.3, pp. 397–404. ISSN: 0262-8856. DOI: 10.1016/j.imavis.2007.01.006. URL: http://dx.doi.org/10.1016/j.imavis.2007.01.006 (cit. on p. 14).

Vu, H. H. (2011). "Large-scale and high-quality multi-view stereo." PhD thesis. Université Paris-Est (cit. on pp. 4, 6, 12, 13).

Zach, C., T. Pock, and H. Bischof (2007). "A Globally Optimal Algorithm for Robust TV-L1 Range Image Integration." In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 1–8. DOI: 10.1109/ICCV.2007.4408983 (cit. on pp. 14–16).