PhD Thesis

Semantic Interpretation of Digital Aerial Images Utilizing Redundancy, Appearance and 3D Information

Stefan Kluckner

Graz University of Technology
Institute for Computer Graphics and Vision

Thesis supervisors
Prof. Dr. Horst Bischof
Prof. Dr.-Ing. Wolfgang Förstner

Graz, February 2011

Our heads are round so our thoughts can change direction.

Francis Picabia

Statutory Declaration

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 marked all material which has been quoted either literally or by content from the used sources.

Place	Date	Signature

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

Ort	Datum	Unterschrift

Acknowledgement

First of all, I would like to express my deepest gratitude to my advisor Horst Bischof, who permitted me to pursue a PhD at the Institute for Computer Graphics and Vision. Thank you for giving me the opportunity to experiment in the exciting field of computer vision and for your support and advice during my years at the ICG. I am also grateful to Franz Leberl for providing me with innovative and optimistic ideas and for inspiring me to publish at photogrammetric conferences. Many thanks to my second adviser, Professor Wolfgang Förstner for his valuable and constructive comments at the final step of this thesis. Moreover, I would like to thank the people working at Microsoft Photogrammetry: Joachim Bauer, Michael Grabner, Konrad Karner and Barbara Gruber-Geymayer, for their help on how to manage the enormous amount of aerial data successfully. I would like to express my appreciation to my colleagues from the ICG's vision group. Especially the meetings of the Virtual Habitat group have led to many fruitful discussions about further research directions. I want to thank Tom M., Mike, Helmut, Markus U., Martin U. and Thuy for sharing ideas and writing one or two papers together with me. A special thank goes to Hayko for being my conference room mate and for proof-reading and another special thank you to Peter and Tom P. for their help and patience in answering my many questions. Furthermore, I would like to thank the administrative staff of the ICG Renate, Christina, Manuela, Karin and Andi. I am grateful to my colleagues and friends Arnold, Manni, Philipp and Suri, which whom I spent many amusing coffee breaks, non-scientific lunch hours in the park and long after-work sessions. Above all, I am very grateful to all my friends and my family - my father, mother, sister and my grand parents, even though they still do not understand what my work is all about. Most importantly, I would like to thank Claudia for her love, support and patience during busy times of completing this thesis.

This thesis was created at the Institute for Computer Graphics and Vision at Graz University of Technology between 2007 and 2010. During this time, the work was supported by the Austrian FFG project APAFA (813397) under the FIT-IT program and the Austrian Science fund under the doctoral program Confluence of Vision and Graphics W1209.

Abstract

One of the fundamental problems in the area of digital image processing is the automated and detailed understanding of image contents. This task of automated image understanding covers the explanation of images through assigning a semantic class label to mapped objects, as well as the determination of their extents, locations and relationships. Occlusions, changes of illumination, viewpoint and scale, natural or man-made structures, shape-defined objects or formless, homogeneous and highly-textured materials complicate the task of semantic image interpretation. While purely appearance-driven approaches obtain reliable interpretation results on benchmark datasets, there is a trend toward a compact integration of visual cues and 3D scene geometry. Accurately estimating 3D structure from images has already reached a state of maturity, mainly due to the cheap acquisition of highly redundant data and parallel hardware. Particularly, scene information taken from multiple viewpoints results in dense 3D structures that describe the 3D shapes of the mapped objects. This thesis therefore presents methods that utilize available appearance and 3D information extensively. In the context of highly-redundant digital aerial imagery we derive a holistic description for urban environments from color and available range images. A novel statistical feature representation and efficient multi-class learners offer the mapping of compactly combined congruent cues - composed of color, texture and elevation measurements – to probabilistic object class assignments for every pixel. On the basis of the derived holistic scene description, variational fusion steps integrate highly redundant observations to form high-quality results for each modality in an orthographic view. We finally demonstrate that the holistic description, which combines color, surface models and semantic classification, can be used to construct interpreted large-scale building models that are described by only a few parameters. In the experimental evalutation we examine the results with respect to available redundancy, correctly assigned object classes, as well as to obtained noise suppression. For the evaluation we use real-world aerial imagery, showing urban environments of Dallas, Graz and San Francisco, besides standard benchmark datasets.

Kurzfassung

Eines der grundlegenden Probleme im Bereich der digitalen Bildverarbeitung ist das automatische und detaillierte Verstehen von Bildinhalten. Die Aufgabe der automatischen Erklärung von Bildern umfasst die Interpretation der Bilder durch die Zuweisung einer semantischen Bezeichnung abgebildeter Objekte, sowie die Bestimmung von deren Ausbreitung, Positionen und Beziehungen. Verdeckungen, wechselnde Beleuchtung, Blickrichtungen und Größen, natürliche oder vom Menschen geschaffene Strukturen, formdefinierte Objekte oder formlose, homogene und strukturierte Materialien erschweren die Aufgabe der semantischen Interpretation von Bildern. Während rein visuell-basierte Ansätze zuverlässige Ergebnisse auf Vergleichsdatensätzen erzielen, ist ein Trend zur kompakten Integration von visueller Information und 3D Szenen Geometrie erkennbar. Die exakte Schätzung der 3D Struktur von Bildern hat bereits Marktreife erreicht, vorallem aufgrund der kostengünstigen Generierung von hoch-redundanten Daten und erschwinglicher Hardware mit Parallelberechnungsunterstützung. Besonders Szenen Information, aufgenommen aus mehreren Blickwinkeln, ermöglicht die Berechnung vollständiger 3D-Strukturen, welche die 3D-Formen der abgebildeten Objekte beschreiben. Diese Arbeit behandelt Methoden, die das vorhandene Erscheinungsbild, sowie 3D-Information umfassend ausnützen. Im Kontext von hoch-redundanten digitalen Luftbildern präsentieren wir Ansätze für die Berechnung einer ganzheitlichen Beschreibung für urbane Lebensräume aus Farbe und Tiefenbildern. Eine neue statistische Merkmalsrepräsentation und effiziente Klassifikatoren für mehrere Objektklassen bieten eine schnelle Projektion kongruenter Messungen von Farbe, Textur und Höhenwerten zu probabilistischen Klassenzuweisungen für jeden Bildpunkt. Aufgrund der ganzheitlichen Szenenbeschreibung können redundante Beobachtungen mittels Variationsmethoden zu qualitativ hochwertigen Ergebnissen für jede Modalität in einer orthographischen Ansicht fusioniert werden. Wir zeigen abschließend, dass diese ganzheitliche Beschreibung, bestehend aus Farbe, Oberflächenmodell und semantischer Interpretation, für die großflächige Berechnung von Gebäudemodellen mit wenigen Parametereinstellungen herangezogen werden

kann. Mittels synthetisch erzeugter Modelle untersuchen wir in Experimenten die erzielten Ergebnisse im Hinblick auf vorliegender Redundanz, korrekt zugeordneter Objektklassen, sowie erzielter Rauschunterdrückung. Für die Auswertung nutzen wir neben Vergleichsdatensätzen, reale Luftbildaufnahmen von Dallas, Graz und San Francisco.

Contents

Chapter 1

Introduction

Three-dimensional reconstruction and object recognition are two major trends, that dominate the current research in both computer vision and photogrammetry. The automatic estimation of the geometry of arbitrary image scenes has already reached a state of maturity, due to highly redundant image data and increased computational power through multi-core systems. They enable stereo or multi-view matching methods that result in highly accurate 3D scene information, *e.g.*, [Hirschmüller, 2006, Agarwal et al., 2009, Goesele et al., 2010,Frahm et al., 2010,Irschara, 2011]. Although object recognition has been studied as intensively as 3D reconstruction, it is still a wide area of active research. Especially huge inter- and intra-class variabilities, scale and illumination changes and partial occlusions present research with a challenge. The semantic interpretation of high-resolution aerial images is a particularly difficult task, since every pixel must be assigned a semantic class label. In this thesis we face these challenges by arguing, that it is essential to integrate available appearance cues, such as color and texture, and 3D information, computed from overlapping scene observations, for the task of semantic interpretation. Combining congruent color, 3D information and semantic knowledge about an observed scene takes us a giant step toward a full image understanding and holistic description. We therefore propose concepts, that utilize redundant input sources exceedingly and investigate the interaction between appearance and scene geometry to improve the task of aerial image understanding.

1.1 Aerial Photogrammetry

Aerial mapping is important to governmental and commercial organizations all over the world. During the last two decades digital cameras have started to replace traditional analogue cameras and took over the market of aerial photogrammetry, as a result of their enhanced stability and resolution [Leberl et al., 2003]. Figure 1.1 demonstrates the quality

Figure 1.1: Small image patches cropped from huge digital aerial images at pixel sizes of 8 cm and 15 cm, respectively. One can notice that fine image details, such as walking people, vegetation or railway tracks are preserved accurately.

of digital aerial images at pixel sizes of 8 cm to 15 cm. Although these images are taken in flight heights of approximately 1000 m, one can recognize fine details like pedestrians and railway tracks.

Digital aerial photogrammetry today offers various applications, that permit the creation of accurate surface models, as well as the generation of eye-candy orthographic images and the construction of detailed models of urban environments, visualized efficiently on the Internet. For example, location-aware applications, like the initiatives *Microsoft Bing Maps* and *Google Earth*, aim at reconstructing all major cities of the world in order to offer accurate virtual models of real urban environments. They additionally collect meta-data about hot-spots, like touristic sights, shopping malls and churches, as well as residential houses or single trees, and link it to the aerial data to enrich location-aware search engines. Another big advantage of digital aerial mapping lies in the possibilities it offers to Geographic Information Systems (GIS), due to low cost and efficiency. GIS applications typically include urban planning, cartography, emergency management, mapping of roads, railway lines, fiber-optic corridors or pipelines, and navigation. Especially car navigation has become ubiquitous, since it facilitates driving in foreign areas. To model entire urban environments or link aerial data to existing GIS and navigation systems automatic procedures and autonomous interpretations are needed, in order to handle the enormous amount of collected information and to understand the mapped data. These applications thus depend on an interpretation of urban scenes, which identifies roads,

buildings, trees, doors, windows, parking meters, sky lights and similar objects.

1.2 Photo-realistic Modeling vs. Virtual Cities

The use of highly overlapping, digital aerial photography makes the computation of 3D scene information possible [Hartley and Zisserman, 2000, Kraus, 2004]. Purely image-based methods offer intermediate results, like surface and terrain models. While a digital surface model (DSM) includes elevation measurements for buildings, vegetation, and roads, as well as for the visible terrain, the digital terrain models (DTM) represent the bare-earth surface of which elevated objects, like man-made structures and trees, get stripped. Both models are typically represented in the form of sampled elevation measurements. Associated with the points, or triangular meshes, computed from the 3D elevation measurements, are small patches of photo-texture information directly extracted from the digital aerial images. From these intermediate results, one builds both high-quality photo-textures as well as complete 3D models of urban environments. Although two-dimensional ortho-photos provide a simpler data structure, smaller data volume and a great ease of use and orientation, they do not inspire the observer. On the contrary, computed photo-realistic 3D models, consisting of huge point clouds or assemblies of triangular meshes with attached photo-texture indeed please the eye but need complex data structures, and are more difficult to use and navigate. Both representations ortho-photos and photo-realistic 3D models yet do not present any semantic knowledge that is required for navigation, searching, data mining or GIS applications. We thus cannot query the data according to content, nor can we derive specific thematic information of a specific object class.

The ultimate goal must be a full explanation of all mapped objects in the scene by attaching semantic labels to them, so that one can search through these objects or model them according to the obtained interpretation. A virtual city [Nicklas, 2007], the visualization of an urban scene, will then be created from the complete description of its elements. Apart from this flexibility, there is an advantage of smaller data quantities that need to be stored, retrieved and transmitted to an user. Creating the rendering from object models additionally supports the idea of user-supportive generalization. Objects of importance, like shopping malls, restaurants, leisure areas etc., might be visualized at to the user at a level of detail (LoD) and scale, that differs from other objects of less importance. A German initiative deals with standards based on a mark-up language CityGML[1] [Kolbe et al., 2005] and describes the virtual city by five levels, starting from a LoD-0 defined by the DTM. LoD-1 is the simple rectangular model of buildings and city blocks with-

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

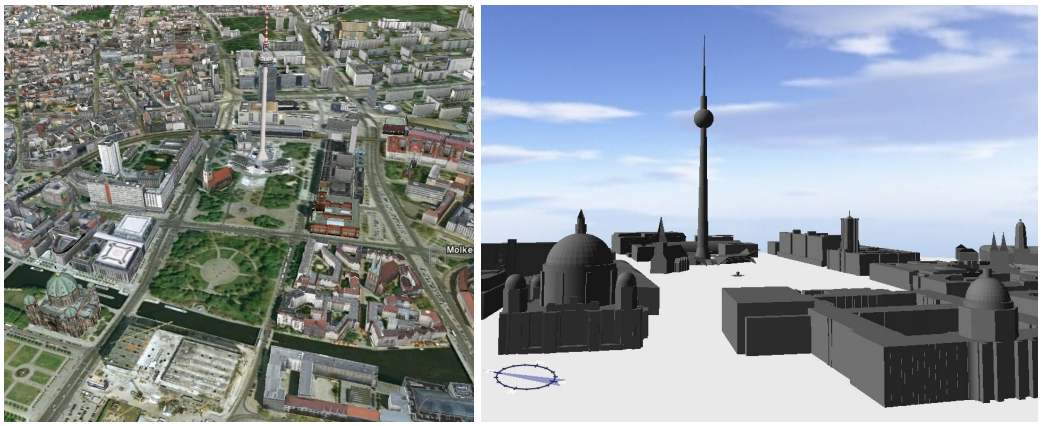


Figure 1.2: A scene of *Berlin* rendered and visualized in **Google Earth** (left) and the corresponding CityGML model (right) with overlaid buildings, represented in LoD 1 to 3. Note that both models have been generated with massive human interaction.

out any attention to photographic realism. This gets improved by a LoD-2 with building blocks showing generalized rooftop shapes. LoD-3 is the photo-realistic full model of each building and LoD-4 contains sufficient details to enter a building. Standardized files, compactly representing the queried data, can be streamed efficiently over the Internet and get rendered immediately at the users workstation. Figure 1.2 depicts a photo-realistic visualization provided by *Google Earth* and a corresponding CityGML rendering of a scene located in *Berlin*. Similar to the CityGML standard, procedural modeling [Parish and Müller, 2001] aims at constructing synthetic 3D models and textures from sets of defined rules and shape grammars.

1.3 Semantic Interpretation of Aerial Images

In order to construct semantically enriched virtual cities that support search queries or specific information extraction, every mapped object has to be recognized and assigned a semantic class label as a first step toward a full image understanding. The problem of image understanding covers the interpretation of the scene with respect to mapped objects, locations and the their relationships. Figure 1.3 shows a scene taken from two different viewpoints. The high overlap enables an estimation of 3D scene geometry, which can be used to improve the semantic labeling of the scene. In this thesis we treat the problem of scene understanding as a semantic labeling process, where each pixel is assigned a specific object class.

Due to huge variability the task of visual scene understanding is still a largely unsolved problem. Among occlusions, illumination, viewpoint and scale changes, natural or man-

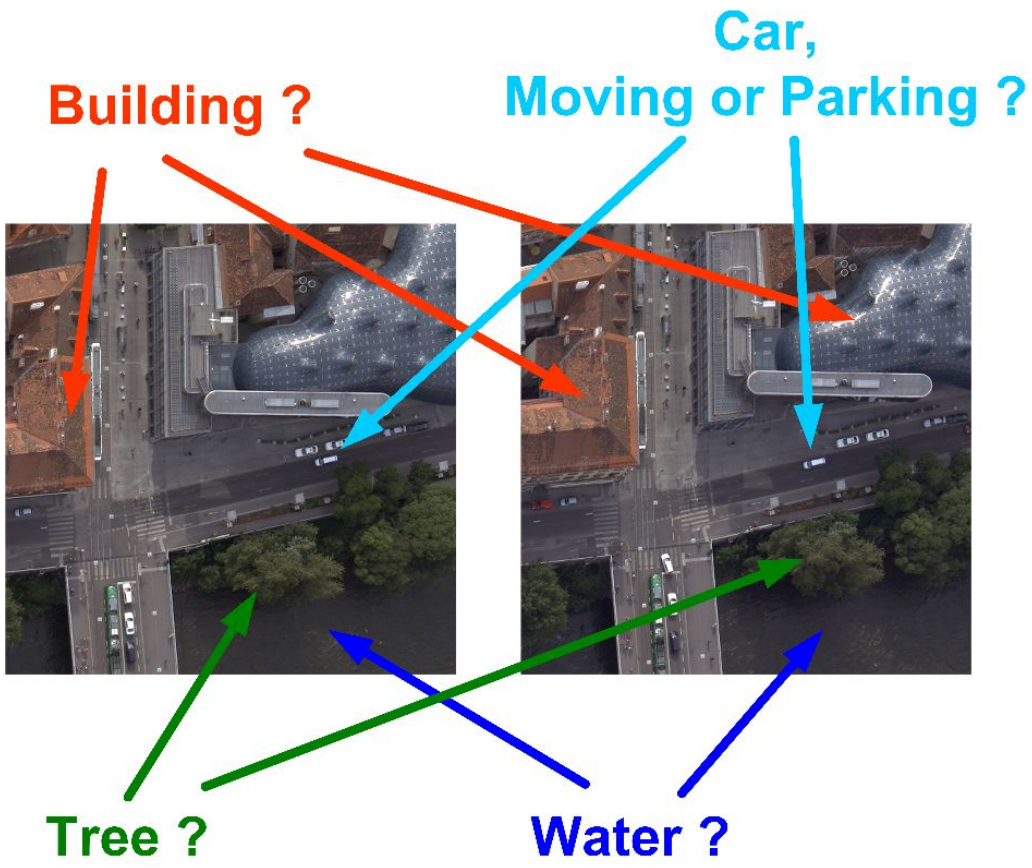


Figure 1.3: Semantic interpretation aims at assigning class labels to mapped objects. In this thesis we focus on a pixel-wise scene understanding by using a given set of predefined object classes. The high redundancy in the aerial data enables an estimation of the scene geometry as well as a semantic classification from different viewpoints.

made structures, shape-defined (*things*) or formless objects (*stuff*), transparent or highly-textured regions complicate the task of finding a meaningful object representation for effective scene understanding. Visual scene understanding has been addressed by a variety of prominent works [Shotton et al., 2007, Rabinovich et al., 2006, Verbeek and Triggs, 2007, Gould et al., 2008, Ladicky et al., 2010a] that aim at an accurate explanation of each pixel in observed image. From these approaches, one can notice that the combination of several intermediate results achieves improved recognition accuracies and is required to obtain a full description of images.

Semantic image interpretation can be seen as possibly highly automatic processes for understanding and analyzing the objects shown in arbitrary images. These processes include object detection, aiming at tagging objects of a particular target class with a bounding box, but also image classification, where pixels, assemblies of pixels, parts or even

images are assigned a label selected from a pool of predefined object classes. Object detection differs from classification in its goal. The goal of object detection is to describe, count and find an object as a whole[1], by taking into account its visual appearance and shape, represented either in 2D or 3D.

We consider image classification as a pixel-oriented process, where a single pixel or a pixel and its neighborhood are the input and a semantic class label for a single pixel is the output. In addition, the task of image classification also treats the relations between pixels within the observed image. The choice of using object classification or detection is mainly based on a discrimination into *things* and *stuff* object classes [Forsyth et al., 1996]. While *things* objects have a distinct size and shape, *stuff* can be rather seen as a material that is specified by a texture and has no ascertainable form[2]. Distinguishing between these two types leads to improvements for the task of full scene description as recently shown in [Heitz and Koller, 2008, Ladicky et al., 2010a].

This thesis mainly focuses on the semantic enrichments, also referred to as land-use classification, of mapped object points, greatly represented with high-resolution and highly redundant aerial imagery. The image content gets separated into regions that typically describe buildings, circulation spaces, such as streets and parking areas or driveways, vegetation, grass surfaces and water bodies. The regions are commonly defined by assemblies of pixels forming an ineffable amount of object variabilities and spatial extends. We thus treat scene understanding as *stuff* classification, where each pixel is described by its local neighborhood [Shotton et al., 2007, Verbeek and Triggs, 2007]. Although there exist solely image-based interpretation approaches for aerial images [Gruber-Geymayer et al., 2005, Porway et al., 2010], none of these concepts make considerable use of a tight combination of appearance and 3D information. We therefore propose to utilize these complementary cues for a semantic enrichment of digital aerial images. For instance, using a direct combination of color and height data would successfully separate gray-valued rooftops from streets or might be useful to distinguish between grass and elevated vegetation, like trees and bushes. Having the semantic explanation of each pixel will have a significant effect on many following tasks. We think of the way we search in images, how we can compress, publish and broadcast them on the Internet or how urban scenes get represented and visualized. Moreover, assigning an object class to each object point, brings us, together with available color and 3D information, directly to a holistic description of aerial images scenes. This holistic collection supports the refinement of the obtained interpretation toward a fine-scaled description, where building blocks, rooftop landscapes, chimneys, trees etc. are represented individually.

[1] In detection tasks the object is usually known in principle to have a certain scale, shape or appearance

[2] It is obvious that the assignment of objects to the *things* or *stuff* category depends on the image acquisition, *e.g.*, clouds or trees might then be refer to both categories

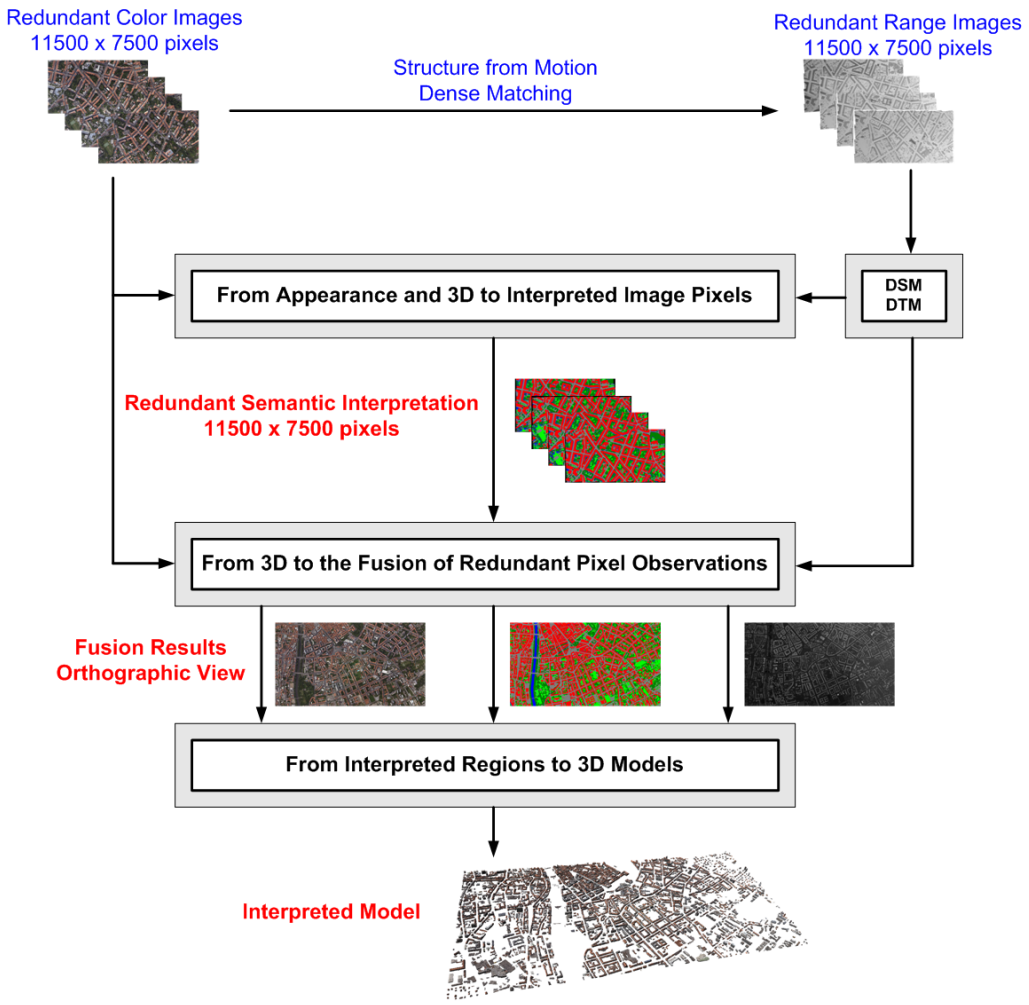


Figure 1.4: Overview of the proposed aerial image interpretation workflow. We utilize highly redundant color and range images for the pixel-wise explanation. A fusion step provides us with a holistic scene description, consisting of fused color, height and classification, that can be used to construct interpreted city models.

1.4 Contributions

In order to obtain an accurate semantic interpretation of aerial images, this thesis focuses on utilizing highly redundant appearance and 3D information cues extensively. In particular, modern digital imaging techniques enable image acquisition with high redundancy from significantly overlapping views without high additional costs. In this thesis we thus study the combination of redundancy, appearance and scene geometry for tightly interacting processing steps. Figure 1.4 summarizes the proposed workflow.

1. From Appearance and 3D to Interpreted Image Pixels The first contribution deals with large-scale semantic interpretation of single high-resolution aerial images. In order to obtain a reliable object class assignment to each pixel in the images, we introduce a novel region descriptor, based on statistical measurements. This descriptor enables a compact and straightforward integration of appearance and 3D information without considerations about normalization issues. Since the aerial imagery provides an enormous amount of data, we exploit efficient data structures, such as integral images and fast multi-class learners. To provide a quick descriptor calculation, the proposed supervised classification approach mainly relies on a pixel-wise semantic interpretation by incorporating local neighborhoods. We also exploit unsupervised segmentation, providing important spatial support, and optimization strategies to accurately delineate the mapped objects with respect to the real edges. The suggested interpretation workflow gets first evaluated on standard benchmark datasets. For the aerial images we then demonstrate that the tight combination of appearance and 3D information is essential for an accurate semantic interpretation at the pixel level.

2. From 3D to the Fusion of Redundant Pixel Observations. The second contribution studies the fusion of redundant observations for the semantic interpretation, color and height measurements. The generation of these congruent modalities is essential for a holistic scene description. Due to the high geometric redundancy of the aerial imagery and the availability of well-defined scene geometry, a projection to a common view (we employ the orthographic view) yields multiple, highly redundant pixel observations for each involved modality. We thus make use of the range images, representing the geometry, to align the overlapping perspective image information into a single view. Since the range images are prone to contain many outliers we put emphasis on optimization schemes, offering novel regularization techniques. As a first step we exploit continuous energy minimization methods to integrate redundant scene observations for color and height. The fusion step, defined over multiple color observations, utilizes improved image priors, based on efficient wavelets. In contrast to the height field fusion, where a total-variation based regularization advances sharp edges, the wavelet regularization preserves fine structures and a natural appearance. A second step yields congruent semantic interpretation within the orthographic representation by accumulating multi-view classifications, resulting from the per-image explanation. Having the raw ortho-fused interpretation facilitates an investigation of different optimization methods with respect to correctly classified pixels and object delineation. The experiments show that the continuous formulation of the Potts model provides a simple way to capture object edges by incorporating appearance or height data within the optimization procedure.

3. From Interpreted Regions to 3D Models. The third contribution addresses the modeling of the obtained semantic interpretation. We particularly discuss the construction of fully parametrized 3D models of buildings. The proposed approach again relies on the tight interaction between semantic classification, color and height information. In a first step, we make use of unsupervised color image segmentation to extract and model any type of building footprint. Established building regions, derived from the semantic interpretation, are then assigned with individual geometric primitives efficiently estimated from the corresponding height data. A clustering yields representative geometric prototypes that are used within a final labeling procedure. This labeling strategy, defined on a reduced adjacency graph, yields a consistent and smooth modeling of the individual rooftops, where each rooftop gets parametrized by few parameters. Available elevation measurements and a simple triangulation are subsequently used to extrude the final 3D building model, efficiently described by unique ID-number, an assembly of polygons, geometric prototypes and mean colors.

1.5 Outline

We detail the contributions in three technical chapters that describe the semantic interpretation of single images, the fusion of available modalities and the modeling of the obtained interpretation. For each of the main chapters, we discuss related work, outline the concepts and evaluate the methods individually. The remaining part of the thesis is thus organized as follows:

Chapter 2. In Chapter 2 we briefly describe different input modalities required for the proposed algorithms. We present the basic information of the involved aerial imagery, distinguish between radiometric and geometric redundancy and explain the generation of surface and terrain models.

Chapter 3. Chapter 3 deals with the semantic interpretation of single images by tightly integrating appearance and height information. This chapter is divided into four parts: feature representation, feature classification, the application of unsupervised segmentation for improved object delineation and the refinement step based on energy minimization. The experimental evaluation studies the performance of the proposed interpretation workflow applied to standard image collections and aerial images.

Chapter 4. Chapter 4 focuses on the fusion of massively redundant image information. While in Chapter 3 the aerial images are treated independently, this chapter aims at integrating the data from multiple perspective images. We thus highlight the generation of

high-quality results from multiple observation for color, height and semantic interpretation. For each modality separately, optimization schemes are introduced for the fusion in an orthographic view. The evaluation section demonstrates the benefit of using the high redundancy provided by the aerial imagery.

Chapter 5. In Chapter 5 we address the problem of semantic modeling by incorporating the holistic scene description obtained from Chapter 4. The main part of Chapter 5 covers the extraction, parametrization and modeling of rooftop landscapes from color information, derived building regions and surface models. A detailed evaluation shows the benefit of this interaction.

Chapter 6. Finally, Chapter 6 concludes the thesis with a summary of the obtained outcomes and gives and ideas for future work, that could help to improve the suggested interpretation workflow.

Chapter 2

Digital Aerial Imagery

This thesis aims for a largely automatic semantic interpretation of digitally mapped urban environments by utilizing massively redundant information of appearance and 3D scene information. While highly overlapping color images and corresponding geometric data are assumed to be given in advance, additional products are required to achieve the interpretation of aerial scenes. We thus outline involved modalities, ranging from redundant color and texture information, over models, representing the surface and the terrain, to estimated elevation measurements. Having the holistic collection of data applies to semantically describe entire virtual space, but also to automatically derive parameters individually specified for a certain task. In this thesis the proposed approaches largely use digital aerial images produced by the high-resolution camera *Microsoft UltraCamD* [Leberl et al., 2003]. The camera provides multi-view pixel observations, obtained by acquiring the aerial project with high overlaps, as well as multi-spectral information for mapped point, so that any applied algorithm can rely on highly redundant analysis. We first describe the digital aerial imagery and highlight the process of range image generation. This chapter further outlines the generation of additional input sources, such as the terrain model and the derivation of the elevation measurements, and discusses the orthographic image representation that is commonly used in aerial workflows. In addition, we introduce the aerial project *Dallas*, *Graz* and *San Francisco*.

2.1 Aerial Photography

Digital aerial mapping has become feasible with the first large-format digital aerial cameras in the beginning of the 21st century. To compete with at that time traditionally used film images, these digital cameras had to offer an image width of at least of 11000 pixels and a color band including RGB and infrared channels [Leberl et al., 2003]. The first dig-

ital camera[1] in use was inspired from space cameras in the form of a push-broom sensor operating with a linear array that images a single line in the terrain at any one time. Alternative products[2] were frame-based, but operating a rectangular instead of a square format. A large frame format is being achieved by combining image tiles obtained through separate lenses. An additional factor is the color which in novel frame systems is obtained at a lower geometric resolution than the panchromatic channel, and the final image is a result of fusing panchromatic with color bands. The issue of forward motion compensation is implemented in some of the camera types to achieve longer exposure times and thus better radiometry. Current large-format aerial cameras include multiple area charge coupled devices (CCD) based on the Bayer-pattern color and up to eight lenses [Leberl et al., 2003, Gruber, 2007]. They thus provide an increased stability and offer geometric resolutions in the range of 195 MPixels at a radiometric resolution of 13 bits per pixel. A typical pixel size (ground sampling distance (GSD)) of today's cameras is at 8 to 15 cm. Figure 2.1 shows several detailed views of *Graz* demonstrating pixel sizes of 8 cm.

On the contrary, digital satellite imagery offers pixel sizes in the 50 cm range and are very useful to create coarse ortho-photos, but the computation of interpreted images and the creation of accurate 3D models are first being obstructed by an insufficient resolution, and secondly by an insufficient geometric redundancy (mainly suppressed by a projection near to the orthographic ones). Digital aerial sensing therefore offers (i) a dramatic improvement of image quality over film images from then perhaps 7 to 8 bits per pixel to today's 13 bits, (ii) the no-cost option of increasing the number of images per project and thus vastly improving the redundancy of the taken measurements, and (ii) the ability of increasing the level of detail by reducing the pixel size in a project at almost no additional cost. The combination of these improvements in image quality and image overlaps has favorable consequences for the degree of automation for further processing step on standard workstations or multi-core systems.

2.2 Redundancy

Digital aerial sensors today produce highly redundant digital aerial images cost-efficiently. With traditional analogue film images, a 150 sqkm urban space in the past was imaged from an aircraft with a simple two-image overlap at 60% in the direction of flight and 20% across the flight direction. While the traditional film photogrammetry maps each object point on only a stereo pair, a digital system can increase the geometric redundancy to ten or even more without additional costs. For digital images, one can show that the increase from two to ten or more observations per object point improves the measurement

[1] www.leica-geosystems.com
[2] www.intergraph.com/photo

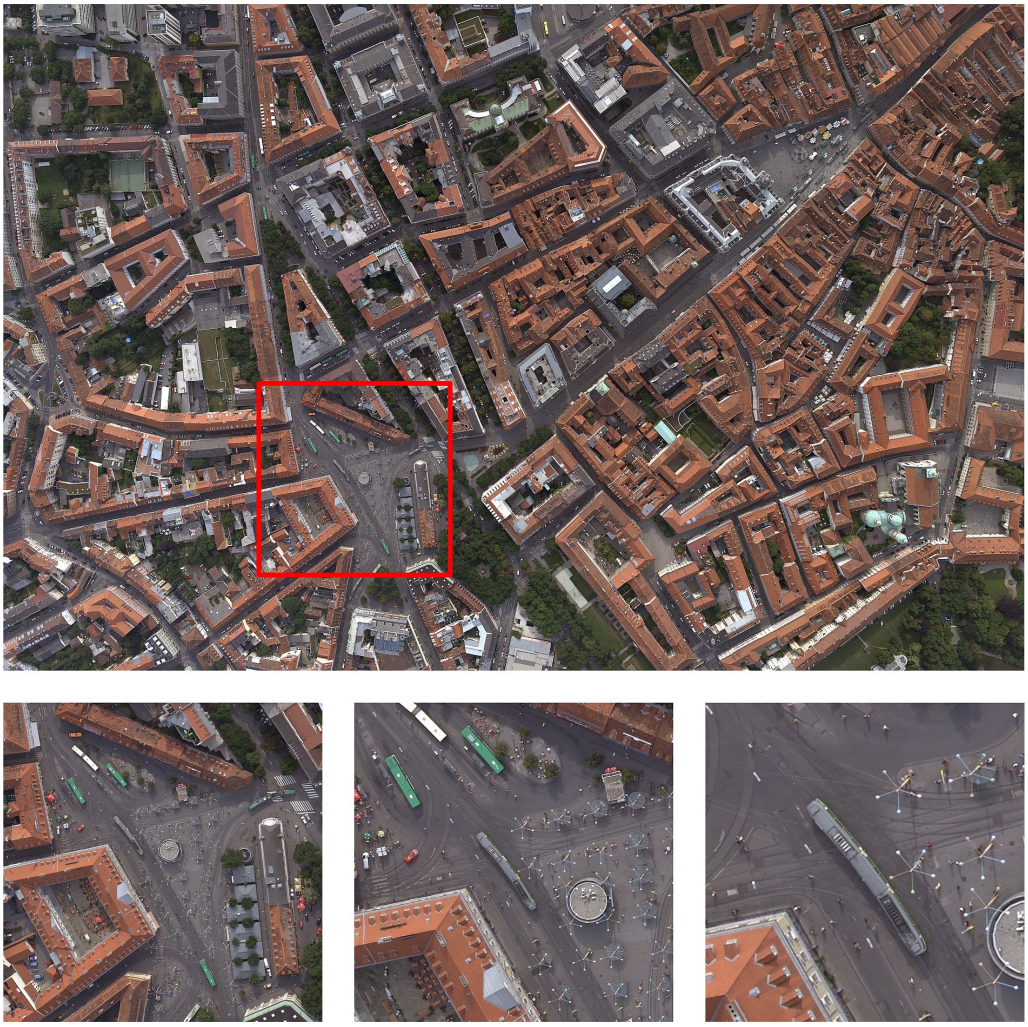


Figure 2.1: Geometric resolution and pixel size. The first row depicts a single image taken from the dataset *Graz* with a geometric resolution of 11500×7500 pixels. Each pixel describes an area of approximately 8×8 cm. The second row shows details cropped from the huge image. Note that at this pixel size walking humans can be recognized.

accuracy by a factor of six [Ladstätter and Gruber, 2008]. It also reduces the occlusions between high buildings and it supports mapping of facades since they will become visible near the edges of vertical aerial images. The increased geometric accuracy particularly enables measurements with improved precision, which is needed to obtain sharply delineated object edges, *e.g.*, in the dense matching procedure (see Section 2.3) or a multi-view semantic interpretation workflow (see Chapter 3 and 4).

The pixel sizes obtained by scanning analogues films may have been at 25 cm on the ground and offer RGB color channels at 24 bits, representing a data volume of 300 MBytes per aerial photograph. A 150 sqkm area today is being mapped on 3000 digital

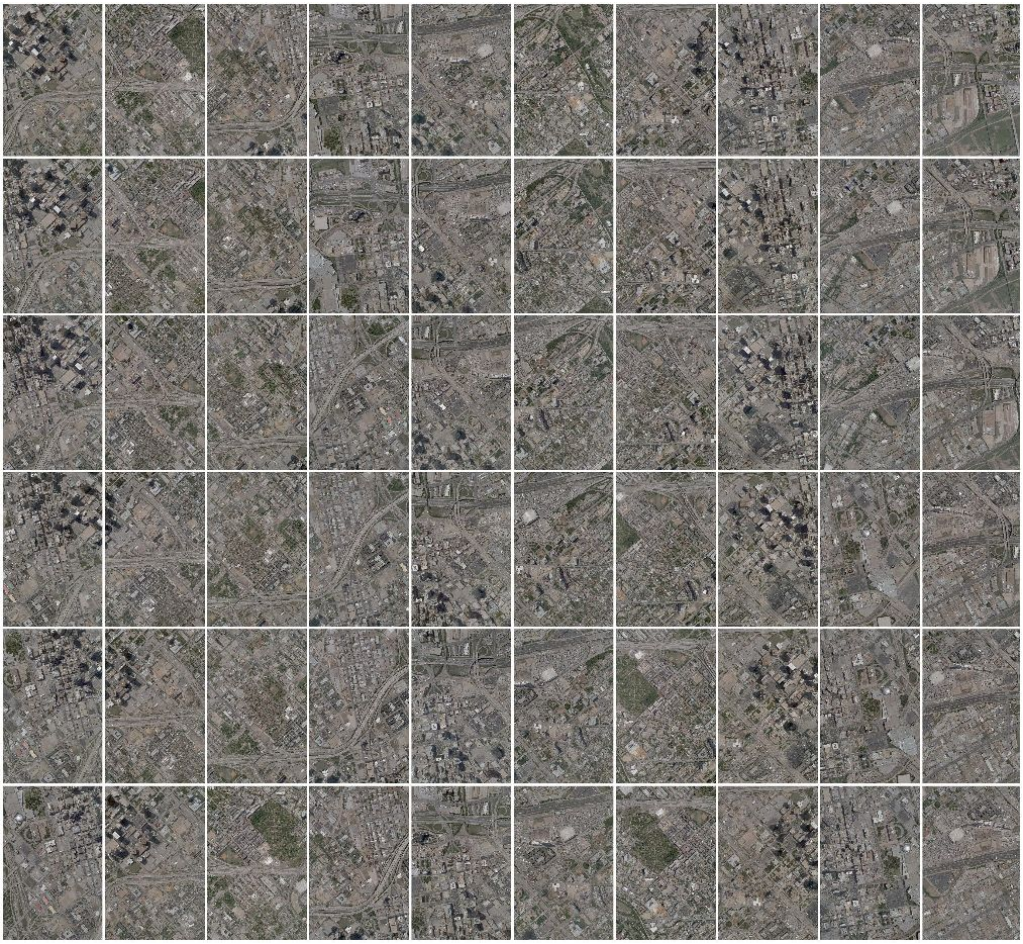


Figure 2.2: Sixty images taken from the dataset *Dallas*. The scene is taken from highly overlapping viewpoints. In this work, we utilize the geometric redundancy to produced a holistic description of urban scenes.

images (16 bits color) with $80\%/60\%$ overlaps and a pixel size of 12.5 cm, resulting a data volume of approximately 1.5 GBytes. This is an increase of the data quantity by two orders of magnitude and is at the core of achieving full automation of aerial mapping. Figure 2.2 depicts sixty highly overlapping images taken from *Dallas*.

The high geometric redundancy is obtained by using forward and side overlaps, so that every point in the terrain is imaged at least 10 times, and any algorithm can rely on multiple analysis results that then can either reinforce or cancel one another. Figure 2.3 shows a set of redundant scene observations taken from *Graz*. Every point on ground is mapped multiple times from different views. The proposed approaches make use of the high redundancy.

The geometric redundancy additionally gets augmented by a radiometric redundancy

Figure 2.3: Some redundant observations of a scene located in *Graz*. The mapped area is shown from different camera viewpoints, which are defined by the flight direction. In our approach we integrate overlapping information within a common orthographic view.

using four spectral bands, adding an infrared band to the classical red, green and blue color channels. Moreover, the high-resolution color images offer sophisticated processing steps that synthetically increase the redundancy. Many essential results, obtained from digital aerial images, like DSM or DTM, classification maps, ortho-photos of any modality and 3D building models, are therefore derived cost-efficiently by fully automated[1] processing pipelines.

2.3 Digital Surface Model

Estimating depth information and hence 3D structure of arbitrary scenes from overlapping views is essential and therefore a well-studied problem [Hartley and Zisserman, 2000, Kraus, 2004]. In this thesis we focus on purely image-based approaches, where we extensively make use of available multi-view matching results computed from the highly overlapping digital aerial images.

[1] Note that the term fully automatic refers to an absence of manual interventions possibly required in case of missing data or erroneous intermediate results.

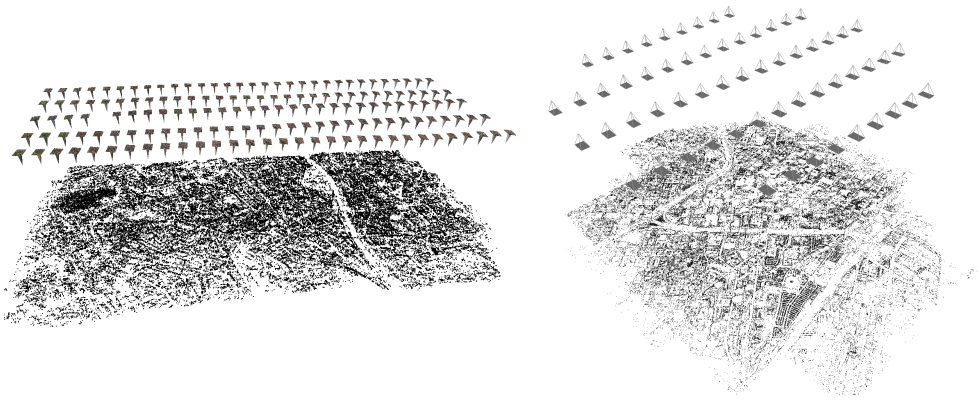


Figure 2.4: Sparse 3D structure and camera orientations computed for the datasets *Graz* and *Dallas* by using the SfM approach described in [Irschara, 2011].

Depth information, represented in the so-called range images, of an observed scene can be computed from at least two overlapping views, generally denoted as stereo image pair. This technique basically determines corresponding points in images (and thus depth information) using, *e.g.*, correlation-based matching methods [Hartley and Zisserman, 2000]. Due to poorly-textured areas and repetitive structures in the images, such simple methods could produce unstable results. It is obvious that the quality of the estimated 3D scene structure will greatly improve when relying on multiple image views. Fortunately, digital aerial images taken with the high-resolution camera, provide a high degree of geometric overlap at no additional cost. Each point on ground is thus visible in multiple images, enabling an improved estimation of the underlying scene geometry.

Dense scene geometry estimation is commonly performed within two processing steps. First, Structure from Motion (SfM) also referred to as aerial triangulation (AT) in the field of photogrammetry, recovers the camera parameters and a sparse 3D point cloud from input images that describes determined pixel correspondences [Kraus, 2004]. As the second step the dense matching estimates depth information for every pixel in the image space [Scharstein and Szeliski, 2001]. In the following section we briefly discuss the principles of SfM and dense matching.

Structure from Motion

The first processing step needed for dense 3D reconstructions, the SfM, recovers the sparse 3D structure of an observed scene and estimates the orientations of the involved cameras. The process of triangulation consists of the manual or automated selection of point measurements in overlapping images (tie points and ground control points) and the determination of unknown parameters of the camera pose and position.

Since the exterior camera orientations are approximately known for an aerial project due to Global Positioning Systems (GPS) and Inertial Measuring Units (IMU), the AT can be rather seen as refinement step in order to obtain camera orientations with sub-pixel accuracy. AT is commonly based on using bundle adjustment [Brown, 1976], which simultaneously refines the scene geometry as well as the estimated camera parameters. The exterior orientations of the aerial images could be achieved by an automatic process as described in [Zebedin et al., 2006], but also by a sophisticated SfM approach as described in [Irschara, 2011], where the camera poses are estimated fully automatically. Figure 2.4 shows the estimated camera poses and the sparse 3D geometry computed for both aerial projects *Graz* and *Dallas*.

Dense Matching

The AT yields camera orientations and a sparse set of 3D points of the scene, but for many practical applications, like a pixel-wise semantic interpretation or the construction of photo-realistic 3D models, dense surface models are highly desired for further processing. A dense matching process therefore estimates depth values for every sampled point in a defined discrete image space. There exists a variety of methods addressing the problem of dense matching problem. A good survey of existing stereo matching methods is given in [Scharstein and Szeliski, 2001]. Depending on the set of pixels on which the optimization of the disparity values is performed, dense matching methods can be roughly categorized into three classes. Among them are local methods [Birchfield and Tomasi, 1998, Yoon and Kweon, 2006], semi-global [Hirschmüller, 2006, Klaus et al., 2006] and global optimization methods [Pock et al., 2008]. Many dense stereo matching algorithms are based on plane-sweep techniques [Cox et al., 1996, Gallup et al., 2007]. The plane sweep concept also allows the direct accumulation of matching costs for multiple images [Hirschmüller and Scharstein, 2009].

By using the recovered camera orientation parameters, an area-based matching algorithm produces a dense range image for each image in the aerial dataset. In our case the range images are computed from a triplet of input images (a reference image and its two immediate neighbors) with the plane sweeping approach described in [Cox et al., 1996]. Figure 2.5 shows three overlapping input images and the corresponding dense matching result. The plane sweeping is based on the normalized cross-correlation as similarity measurement and produces a 3D depth space, which contains the depth hypotheses and their associated correlation values. In order to obtain consistent depth images, where neighboring pixels have similar depth values, the final range images get refined by applying a semi-global optimization approach as described in [Klaus et al., 2006].

In Figure 2.6 some dense matching results are shown for single images extracted from the aerial projects. The computed range images, each with a dimension of 11500×7500

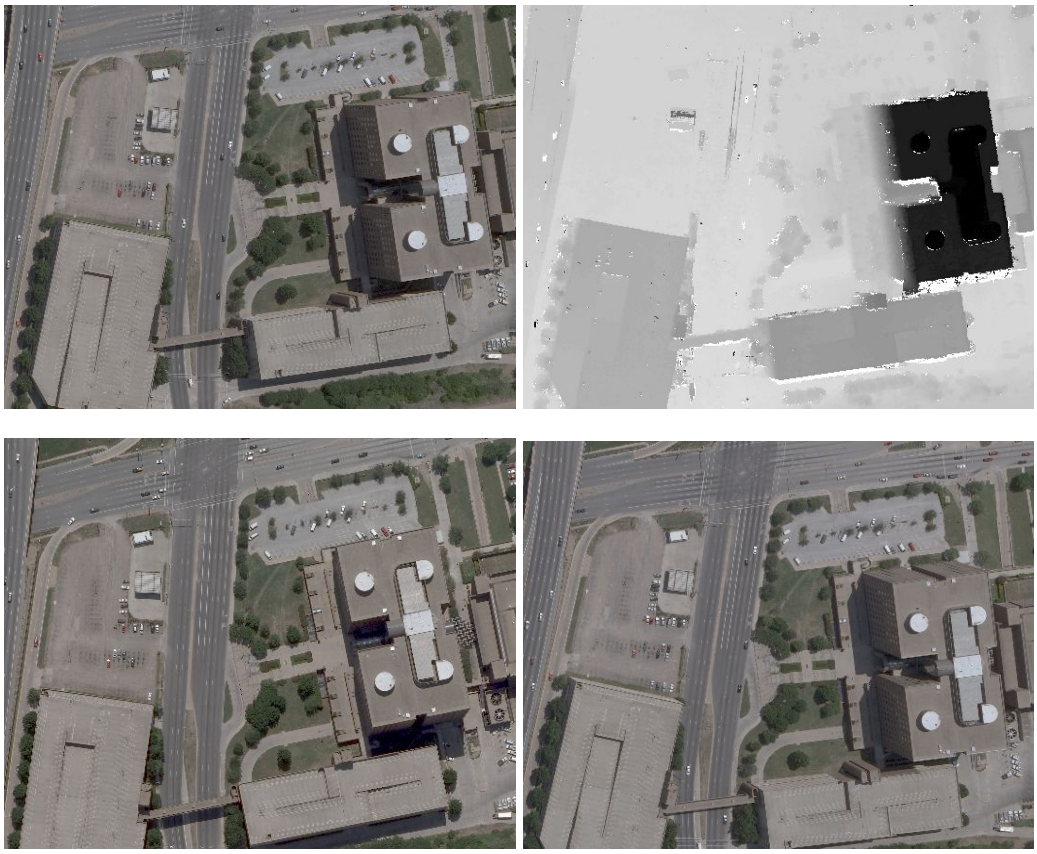


Figure 2.5: A dense matching result obtained for a small scene of *Dallas*. A triplet of overlapping images is used to compute the corresponding range image. The first row shows the center image and the corresponding computed range image. The images in the second row are the direct neighbors, that are taken from different view points. The computed range images are composed of depth values, describing the estimated distances from the camera to the point on the surface of the observed scene for each pixel. These depth values thus define the underlying 3D scene geometry.

pixels, provide depth information for every position in the image space. Each depth value then represents the distance from the camera to the pixel on the estimated surface of the observed scene. Having the depth information for each sampled image point thus enables a pixel-wise transformation to a 3D world coordinate system forming a surface model. Such models can then be defined in 2.5D or in the full 3D space[1].

Since the range data is computed for each image in the aerial project, the high geometric redundancy offers the generation of multiple measurements for one and the same point

[1] Note that we only focus on a 2.5D representation, however, the 2.5D information is sometimes referred to as 3D height throughout this thesis.

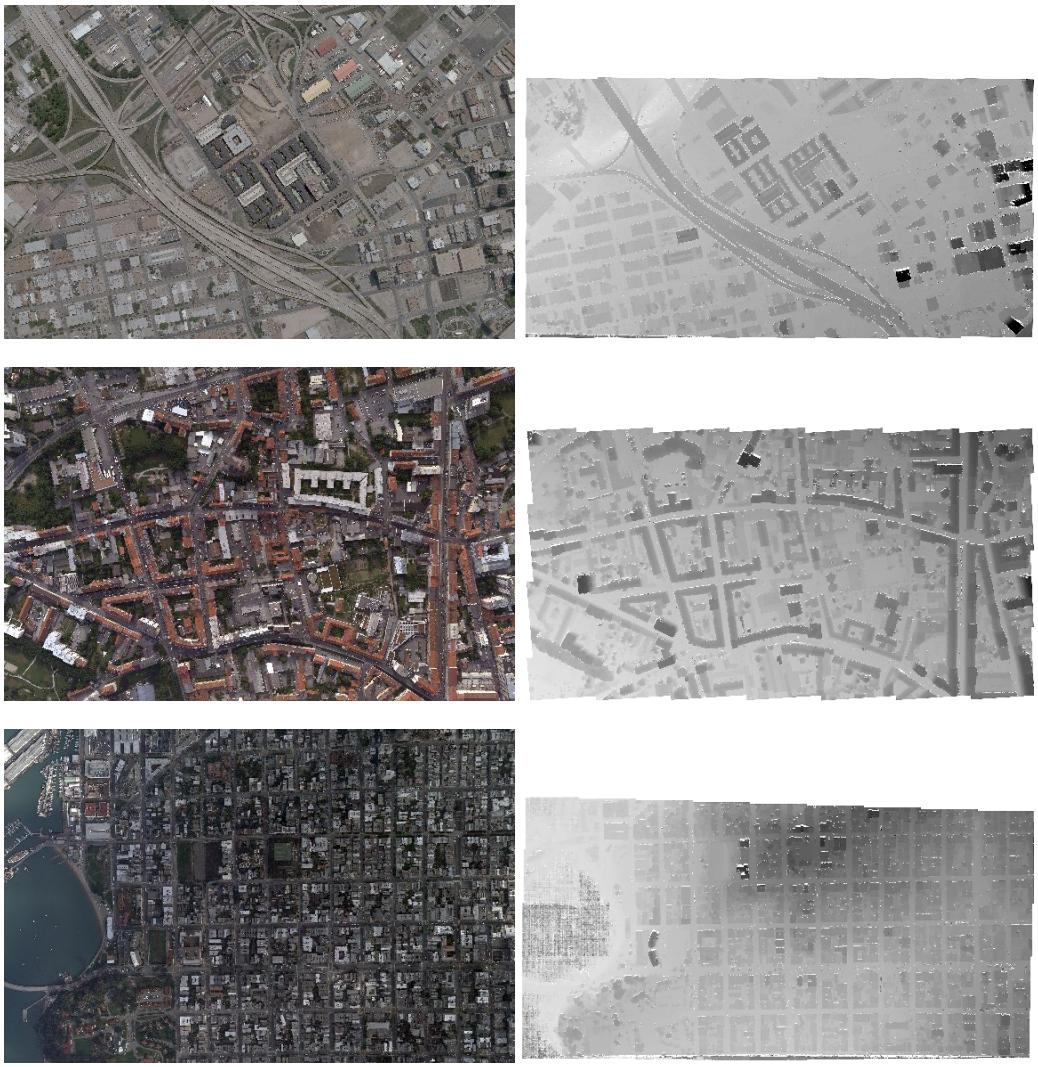


Figure 2.6: Dense matching results, computed for individual images of the aerial projects of *Dallas*, *Graz* and San Francisco. Each range image provide pixel-synchronous depth information for a corresponding color image. By taking into account the camera parameters, depth values can be directly transformed to huge 3D point cloud, forming a sampled surface model in world coordinate system. Depending on the SfM approach, this coordinates system can either be a local or a global one.

of the surface model. Hence, an sophisticated handling of these, highly redundant, 3D observations can be used to improve the surface model itself (as we will show in Chapter 4), but also to derive a terrain model, representing the bald earth. Knowing the terrain model is essential in order to turn relative height observations into a normalized surface model, where the corresponding values denote an elevation from the ground. In the next section

we therefore discuss how to derive the terrain model from the DSM.

2.4 Digital Terrain Model

The dense point cloud, provided by the dense matching process, can be seen as an uninterpreted representation of visible surfaces of observed scenes. In the literature, there exists a variety of techniques to construct a terrain model. These techniques range from applying local operators [Eckstein and Munkelt, 1995, Weidner and Förstner, 1995], over utilizing surface models [Champion and Boldo, 2006, Sohn and Dowman, 2002] and different segmentation [Sithole and Vosselman, 2005] or even recognition approaches [Zebedin et al., 2006], to using hybrid methods [Baillard, 2008].

In this thesis we used a local filtering strategy and a variational strategy to separate the available 3D points into those that describe the terrain surface and those that represent elevated objects. By taking into account the derived range images from overlapping views (we use an approximated median to efficiently integrate the height observations from four neighboring views) and the available camera data, a filtering strategy detects local minimums (defining points on ground) over different scales in the fused point cloud. A derived probability map encodes sampled image regions, that have been formerly elevated by buildings and trees. Similar to the approach described in [Unger et al., 2010], we make use of a variational in-painting strategy in order to fill these areas with meaningful terrain height values.

Subtracting the DTM from the DSM delivers the absolute elevation measurements of the objects, which can now be used as a discriminative feature cue for the proposed semantic interpretation workflow (see Chapter 3) or for the estimation of real object heights, required for the construction of 3D models. Figure 2.7 depicts derived elevation measurements for an image of *Dallas*. Note that these elevation maps provide absolute distance measurements, that can be easily used to distinguish between objects located on the ground and elevated object, like trees and buildings.

2.5 Orthographic Image Representation

In aerial imaging, orthographic projections are a often used to represent the mapped information. Web-driven initiatives, like *Google Earth* and *Microsoft Bing Maps*, frequently offer imagery in the form of 2D ortho-photos due to efficiency. These orthographic images are commonly derived from perspective imagery by transforming the input sources to a parallel projection, where each sampled has then a constant pixel size. Due to the projection, vertical objects, like building facades, are represented as 2D lines in the re-

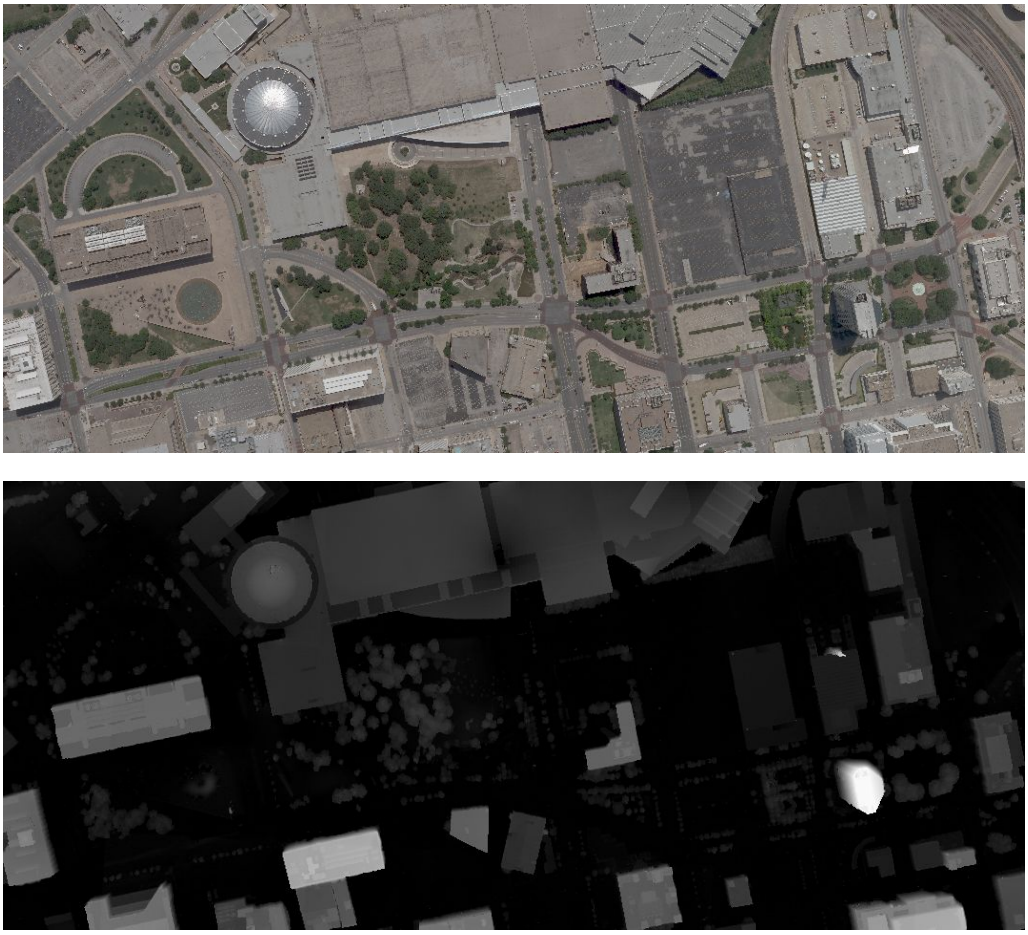


Figure 2.7: Elevation measurements for an image of *Dallas*. The computed elevation maps provide absolute height values for each pixel. Dark values denote no elevation, while bright color defines high elevation. Such normalized surface model are widely adopted for 3D modeling, however, in this work we extensively utilize this representation for the generation of a full semantic interpretation.

sulting ortho-image and, thus, not visible in the resulting view. Figure 2.8 depicts both a perspective and a resulting ortho-projected color image.

The ortho-photos are created by projecting the photo texture either onto a virtual plane for the so-called traditional ortho-photos or onto the DSM to produce a true ortho-image. While proposed methods largely rely on image stitching methods, we extensively concentrate on the redundant data fusion. Knowing the correspondences between image coordinates and the 3D point (resulting from the range images), input sources, like color, infrared, but also pixel-wise semantic knowledge, can be quickly sampled within the orthographic view.

Figure 2.8: An orthographic representation for a *Graz* scene. While the left image shows a perspective view, the right one depicts the ortho-photo of the same scene, where the parallel projection yields constant pixel sizes. It is obvious that vertical structures, like facades, disappear in the ortho-view.

In Chapter 4 we introduce efficient methods for an ortho-image generation providing holistic knowledge about the observed scene, ranging from a fused terrain model, over a high-quality color image, to a pixel-wise semantic explanation of each object point. Hence, a major outcome of this work is a holistic description of aerial data with color, surface models, and semantic interpretation within the orthographic image representation.

2.6 Datasets

In this thesis we apply the proposed methods to three challenging aerial projects, acquired with the *UltracamD* at varying GSDs, namely the imageries *Dallas*, *Graz* and *San Francisco*. The dataset *Dallas* includes large building structures in a relatively flat terrain and is mainly dominated by gray valued areas. The buildings vary from low factory halls to skyscrapers with specular facades. The dataset *San Francisco* has mainly a suburban appearance embedded in a hilly terrain with many building-tree transitions. The 155 images of *Graz* show a colorful characteristic with a detailed rooftop landscape. Different types of vegetation and water areas mapped with a reduced pixel size make this dataset challenging. Table 2.1 summarizes the basic information for the aerial datasets.

Dataset	Nb of Images	GSD	Color, 24 bits	Range images	Infrared
Dallas	68	15 cm	X	X	X
Graz	155	8 cm	X	X	
San Francisco	77	15 cm	X	X	

Table 2.1: The three datasets used throughout this thesis: *Dallas*, *Graz* and *San Francisco*. The images are taken with the *UltracamD* providing an overlap of 80% in the direction of flight and 60% across the flight direction. Each image has a geometric resolution of 11500×7500 pixels. The required data volume per image (used for this work) is in the range of 450 MBytes, including 24 bits color images, the DSM and the DTM.

2.7 Summary

This chapter has described the aerial information sources required for the proposed approaches. These sources range from acquired color information, over derived texture descriptions, to computed surface models describing the geometry of the observed scene. In this thesis we assume that highly overlapping color images, the corresponding range images an the camera data are given in advance. These intermediate results enable an extraction of normalized surface models or the generation of additional appearance cues useful for, *e.g.*, a full semantic image interpretation, but also coordinate transformations from the 2D image space to a 3D world coordinate system. The coordinate transformations particularly allows us to sample multiple observations for color, height and assigned object labels within a common view. Hence, we first outline a concept for a full semantic interpretation of aerial images by compactly utilizing appearance and elevation measurements. Then, the highly overlapping multi-view classification, together with color and geometry, are ortho-projected to form an improved holistic description of urban spaces.

Chapter 3

From Appearance and 3D to Interpreted Image Pixels

In order to explain every visible spot on ground, mapped by the digital aerial imagery, we introduce a novel concept for a full semantic interpretation of urban scenes. Our methods relies on a compact integration of appearance and 3D information cues. In this chapter we show this combination is essential to obtain a reliable semantic description for each image in the aerial dataset. Having an accurate semantic interpretation, as well as both the color and 3D information, then defines a holistic description of mapped aerial scenes. The proposed pixel-wise classification utilizes multiple low-level feature cues, such as true color, edge responses and 3D data, that derived from a variational reconstruction approach. To enable a compact integration of appearance and elevation measurements we thus use powerful region descriptors that are derived from statistical moments. Together with randomized forests probabilistic assignments for multiple object classes are obtained efficiently for each pixel in the image domain. Final processing steps, based on both unsupervised segmentation and energy minimization, additionally refine the assigned class labeling with respect to real object boundaries. The experimental evaluation demonstrates the proposed interpretation method on standard benchmark dataset and compares the obtained rates to state-of-the-art results. We further show that our approach can be successfully applied to two-class problems in aerial imagery, like a building classification, but also to distinguish between multiple object classes like *building*, *street*, *tree*, *grass* and *water*.

3.1 Introduction

Due to indefinitely large variabilities in the object's appearances and shapes, the problem of semantic image interpretation is still an unsolved task in computer vision. Varying

Figure 3.1: Highly overlapping aerial images taken from *Dallas*. The mainly gray-valued areas provide a challenging task to discriminate, *e.g.*, building structures from the street network. We thus argue that 3D information has to be integrated for an accurate semantic scene interpretation.

illumination, partial occlusions, viewpoint and scale changes additionally complicate the problem. Although appearance-driven classification approaches [Shotton et al., 2007, Verbeek and Triggs, 2007, Gould et al., 2008] obtain reliable results on computer vision benchmark datasets, such as the *Microsoft Research Cambridge* (MSRC) [Winn et al., 2005, Shotton et al., 2007] or the *PASCAL* Visual Object Classes (VOC) Challenge [Everingham et al., 2007] images, full scene interpretation in a real world-scale still poses a challenging task. Figure 3.1 shows an aerial scene of *Dallas* that is taken from different viewpoints. Considering the highly overlapping color images, even for humans it is challenging to distinguish between gray-valued building structures and regions that representing the street network. In this work we are certainly not only interested in a reliable explanation of *building* and *street* pixels, but also in accurately separating *tree*, *grass*, and *water* object points. We claim that there is a need to incorporate 3D scene geometry into the classification to permit an accurate semantic explanation of the image content.

In this chapter we thus propose a straightforward yet efficient approach to combine multiple feature cues like color, edge responses and elevation measurements. For instance, using a combination of color and elevation measurements would successfully separate street regions from gray-valued rooftops or would help to distinguish between grass areas and trees. Figure 3.2 shows congruent RGB color and 3D information, extracted from a perspective aerial image of the dataset *Dallas*. A segmentation of these images into several object classes provides then a semantic knowledge of the objects on ground and approves a specified post-processing to construct virtual cities where, *e.g.*, each recognized object is modeled according to its obtained interpretation.

As a first step we outline a rapid semantic interpretation, based on a novel feature representation, that is directly applicable to machine learning methods. The proposed representation, derived from compact covariance descriptors [Tuzel et al., 2006] allows

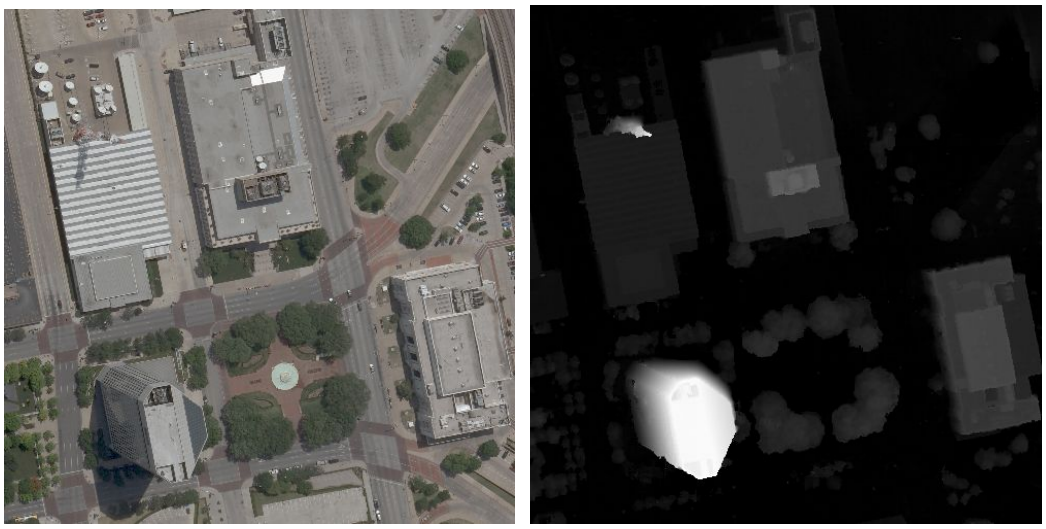


Figure 3.2: Corresponding color and 3D information, represented as a normalized surface model. The proposed concept exploits both appearance and 3D height data to obtain an accurate semantic interpretation at the level of pixels. Bright pixel values denote a high elevation from the ground.

us to compactly integrate color values, derived filter responses or computed height information, while an arbitrary learner efficiently performs a classification task at the pixel level. Due to efficiency and multi-class capability we make extensively use of randomized forest (RF) classifiers [Breiman, 2001], that have been successfully adopted to various computer vision problems. Since the proposed interpretation mainly uses rectangular pixel-neighborhoods we have to consider the accurate object delineation in a second step, where unsupervised segmentation and energy minimization are applied to capture the real boundaries. To demonstrate state-of-the-art performance we present evaluation results for commonly used benchmark datasets by integrating appearance cues. We then apply our proposed method to real world aerial imagery, performing large-scale semantic classification, where the feature representation additionally integrates available 3D information. For both benchmark images and the aerial images the classification accuracy is investigated in terms of correctly classified pixels and by visual inspection.

3.2 Overview

The first processing step of the semantic interpretation workflow involves an initial classification performed on each image in the aerial dataset. We argue that a local *stuff* classification [Forsyth et al., 1996] that is performed at the pixel level, is well suited for efficient

interpretation of images. On the one hand, the aerial images provide a nearly constant *stuff* object scale that is mainly defined by the GSD of the aerial project. The local interpretation can be thus performed for one scale which is mainly defined by the patch size. On the other hand, the huge variability of the mapped data and the undefined object's shapes make a top-down recognition strategy intractable to solve. In addition the training sample generation does not need entirely annotated objects, but rather an efficient assigning of object classes by drawing strokes. We therefore concentrate on a local explanation of the images and introduce the exact object extraction in a later step (see Chapter 5).

We extensively exploit statistical Sigma Points [Julier and Uhlmann, 1996] features, directly derived from the well-established covariance descriptors [Tuzel et al., 2006] in combination with RF classifiers [Breiman, 2001] to compactly describe and classify color, basic texture and elevation measurements within local image regions. The combination of the derived statistical features and RF classifiers provides several advantages for large-scale computations in aerial imagery. Since the aerial imagery consists of multiple information sources, there is a need to reasonably combine these low-level information cues. We therefore apply a Sigma Points feature representation to compactly describe different channels considering a small local neighborhood. Compared to computed histograms over multi-spectral data, these descriptors are low-dimensional and enable a simple integration of appearance and height information, that is then represented on a Euclidean vector space. Moreover, they can be quickly computed for each pixel using integral image structures and also support parallel computation techniques.
Randomized forests have proven to give robust and accurate results for challenging multi-class classification tasks [Lepetit and Fua, 2006, Shotton et al., 2008]. RFs are very efficient at runtime since the final decision is made on fast binary decisions between a small number of selected feature attributes. In addition, the classifier can be efficiently trained on a large amount of data and can handle some errors in the training data.

In the following sections we first review related work in the context of semantic image interpretation and classification. Then, we outline the core parts of our semantic interpretation, consisting of a powerful feature representation (Section 3.4), the classifier (Section 3.5) and refinement steps to obtain an improved final class labeling (see Sections 3.6 and 3.7).

3.3 Related Work

While recently proposed approaches aim at extracting coarse scene geometry directly from interpretation results [Hoiem et al., 2007, Saxena et al., 2008, Gould et al., 2009, Liu et al., 2010] or try to jointly estimate classification and dense reconstruction [Ladicky et al., 2010b], we rather focus on directly integrating available 3D data in our interpre-

tation workflow to improve the task of semantic labeling. The tight integration of 3D data into image classification as additional information source is still a new and upcoming field of research. As shown in [Brostow et al., 2008, Sturgess et al., 2009, Xiao and Quan, 2009], a combination of color and coarse 3D information, obtained from SfM geometry, is essential for an accurate semantic interpretation of street-level images. Leibe *et al.* [Leibe et al., 2007] demonstrated that SfM improves the detection and tracking of moving objects. In this thesis we go one step further by utilizing dense 3D reconstruction. Several approaches in the field of photogrammetry, dealing with aerial imagery, focus on detecting single object classes, *e.g.*, buildings by using only LiDAR data [Matei et al., 2008, Toshev et al., 2010] or height models [Lafarge et al., 2008], but also on exploiting appearance cues together with elevation measurements (resulting from a combination of a surface and a terrain model) [Rottensteiner et al., 2004, Zebedin et al., 2006]. While these approaches focus on binary classification tasks, the presented concept handles multiple classes and can be configured to specific objects. In contrast to [Zebedin et al., 2006], where appearance and geometry are treated separately, our approach tightly integrates dense matching results and low-level cues like color and derived edge responses within a compact yet local feature representation. In addition, we train specified object classes directly and do not introduce prior knowledge, like, *e.g.*, that buildings and trees are elevated from ground, to derive the final classification.

Local classification strategies, using supervision, aim to semantically describe every pixel by considering a small spatial neighborhood or entire image regions, provided by unsupervised segmentation, in the image space. In fact, Bag-of-Features (BoF) models have shown excellent performance in various recognition tasks [Winn et al., 2005, Nowak et al., 2006, Rabinovich et al., 2006, Marszalek and Schmid, 2007, Verbeek and Triggs, 2007, Bosch et al., 2007, Pantofaru et al., 2008, Fulkerson et al., 2009, Lazebnik and Raginsky, 2009]. The BoF concept is mainly based on collecting different types of feature vectors within given image regions. Collected features instances are then quantized into a specified number of words by using well-established clustering procedures. Any image region is then represented by an one-dimensional histogram of word occurrences. The resulting histogram representations get then trained and evaluated with arbitrary classifiers. Commonly, quantized feature instances are extensively composed of combinations of Texton filter bank responses [Winn et al., 2005], SIFTs [Lowe, 2004], Histogram of Oriented Gradients (HOG) [Dalal and Triggs, 2005] or spatial information. However a reliable combination of different feature requires a sophisticated normalization step. An integration of appearance and height information particularly induces problems since height is difficult to quantize, *e.g.*, into one-dimensional histograms if the present value range not known in advance.

Mainly inspired by popular approaches [Viola and Jones, 2004, Tuzel et al., 2007, Shot-

ton et al., 2008], where extracted raw features are directly processed, we build our interpretation step on locally extracted feature descriptors, capable to compactly integrate different cues, which are then trained and evaluated with fast classifiers. Most of the proposed descriptors can be computed rapidly within rectangular image patches by utilizing integral structures [Viola and Jones, 2004, Porikli, 2005, Tuzel et al., 2006] or by directly extracting raw pixel values [Shotton et al., 2008, Gall and Lempitsky, 2009]. In particular, integral images enable descriptor computations in constant time independently of the considered patch size. Viola and Jones [Viola and Jones, 2004] proposed a real-time recognition framework based on Haar-like wavelets. These wavelets are rapidly constructed with intensity integral structure by calculating sums within rectangular image regions. Porikli [Porikli, 2005] proposed to compute histogram-based descriptors with integral structures. This concept was successfully applied to approximate an accelerated construction of region descriptors based on Edge Orientation Histograms (EOH) [Levi and Weiss, 2004] or Local Binary Patterns (LBP) [Ojala et al., 2002]. Prominent techniques like SURF [Bay et al., 2006] and on-line boosting [Grabner and Bischof, 2006] make use of the provided efficiency. Even more simply, Shotton *et al.* [Shotton et al., 2008] developed an extremely fast semantic classification of small images by using RF classifiers [Breiman, 2001] and basic intensity value differences that are evaluated in small spatial neighborhoods. In [Yi et al., 2009] the authors applied an extended approach to the semantic segmentation of medical data that is represented in a 3D voxel space. In our case an extraction of pixel differences might be used to efficiently combine color and 3D height, however, the feature extraction is prone to noisy images (dense matching results usually contain outliers and undefined regions caused by occlusions) and does not provide scale- and rotation-invariance implicitly.

Tuzel, Porikli and Meer [Porikli, 2005, Tuzel et al., 2006] successfully demonstrated the power of various statistical feature representations mainly in the context of visual object tracking. In [Wu and Nevatia, 2008], the authors showed that covariance descriptors can be successfully applied to detection tasks. Especially the covariance descriptors enable a simple integration of various information cues, like intensity values or filter responses and offer an invariance to rotation and scale. Probably the most powerful advantage is the quick generation provided by an extended set of integral images.

Unfortunately, the classification of local feature descriptors, computed within rectangular image regions, does not capture the real object boundaries. In order to accurately delineate the real object boundaries, there is a trend to perform object classification by utilizing conservative over-segmentation techniques. Since our workflow uses a pixel-wise classification by considering a rectangular spatial neighborhood, we propose to introduce an unsupervised segmentation to refine the classification results with respect to object edges. Malisiewics and Efros [Malisiewicz and Efros, 2007] particularly showed that

a correct spatial support, provided by unsupervised segmentation methods, significantly improves the recognition performance over approaches using image patches. A variety of methods, obtaining state-of-the-art performance on evaluation datasets, directly integrates image partition methods into semantic or holistic scene understanding [Russell et al., 2006, Malisiewicz and Efros, 2007, Pantofaru et al., 2008, Kohli et al., 2008, Gould et al., 2009] or object localization and detection [Galleguillos et al., 2008, Fulkerson et al., 2009]. Due to the missing of perfect image partitions [Unnikrishnan et al., 2007], multiple image segmentations are widely used for accurate object delineation and classification [Russell et al., 2006, Pantofaru et al., 2008, Kohli et al., 2008]. It is obvious, that the generation of multiple segmentations induces enormous computational complexity and is impractical for aerial image segmentation. Recently, Fulkerson *et al.* [Fulkerson et al., 2009] proposed to use super-pixels, rapidly generated by Quickshift [Vedaldi and Soatto, 2008], for object segmentation and localization. These super-pixels accurately preserve object boundaries and are thus suitable to describe coherent parts of mapped objects. Additionally, subsequent processing steps, such as a refined labeling, benefit from the reduced image grid in terms of computational complexity [Ren and Malik, 2003]. It is clear that also TurboPixels [Levinshtein et al., 2009] or the small segments provided by [Ren and Malik, 2003] can be used to compute different image partitions.

Once a meaningful, however initial, explanation for a pixel or even for an assembly of pixels is found, image interpretation concepts additionally integrate contextual constraints to find a globally consistent final class labeling. These contextual constraints capture the probability of class occurrences within an image and also provide a regularized labeling in a spatial (pixel or region) neighborhood. Conditional *Markov* random field (CRF) formulations [Boykov et al., 2001, Komodakis and Tziritas, 2007] are widely adopted to include these contextual constraints for coherent image classification and segmentation. The optimization, based on energy minimization, is usually performed on regular image grids [Verbeek and Triggs, 2007, Shotton et al., 2007, Rabinovich et al., 2006, Pantofaru et al., 2008], but also on reduced super-pixel graphs [Fulkerson et al., 2009, Gould et al., 2009].

3.4 Feature Representation

This section highlights the construction of the powerful region-based descriptor, derived from first (mean) and second order statistical moment (covariance matrix). Region covariance representations compactly combine raw pixel values, such as appearance and filter responses, and have shown to give excellent results in object detection [Tuzel et al., 2006, Wu and Nevatia, 2008], visual tracking [Porikli et al., 2006, Tuzel et al., 2007, Arsigny et al., 2007, Tyagi and Davis, 2008, Li et al., 2008] and medical image process-

ing[1] [Pennec et al., 2006, Fletcher and Joshi, 2007]. They additionally provide a lower dimensionality than constructed histograms representations like HOGs [Dalal and Triggs, 2005] or LBPs [Ojala et al., 2002]. Wu and Nevatia [Wu and Nevatia, 2008] and Paisitkriangkrai *et al.* [Paisitkriangkrai et al., 2008] particularly showed that covariance descriptors provide high discriminative power compared to the prominent HOG descriptors.

A major drawback of covariance descriptors is that these matrices do not lie on a Euclidean vector space, which means that element-wise difference computations of two observed matrices do not measure a valid similarity [Pennec et al., 2006]. Therefore, standard machine learning methods, that require similarity computations between attributes or a reference instance cannot be applied directly.

In the following, we first discuss the well-studied first and second order statistical moments. Second, we review related representations based on manifolds and introduce the concept of our novel feature representation that approximates mean and covariances directly on a Euclidean vector space. Additionally, we highlight the efficient construction of the mean and the covariance matrix within rectangular shapes using integral structures, making the descriptor interesting to computer vision problems like large-scale classification or real-time tracking and detection.

3.4.1 Mean and Covariance Descriptors

Estimating the mean and covariance region descriptors from multiple information sources offers a low-dimensional feature representation that simply integrates feature channels, such as color, filter responses, 3D height information, etc., and also exploits the correlation between them. The diagonal elements of the covariance matrix provide the variances of the feature attributes in one channel, whereas the off diagonal elements capture the correlation values between the involved low-level modalities.

In this thesis we assume that the first and second order statistics, represented by the mean vector μ and the covariance matrix Σ, denote the standard parameterization of a D-dimensional, multivariate normal distribution $\mathcal{N}_D(\mu, \Sigma)$. A vector $\mathbf{x} \in \mathbb{R}^D$ is said to be normal distributed according to $\mathbf{x} \sim \mathcal{N}_D(\mu, \Sigma)$. The resulting probability density function is then defined as

$$f_X(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}, \tag{3.1}$$

where $\mu \in \mathbb{R}^D$ and $|\Sigma|$ is the determinant of the covariance matrix Σ. From (3.1) we can observe that the quadratic term in the exponent simplifies to $\mathbf{x}^T\Sigma^{-1}\mathbf{x}$ for zero-mean vectors. Note for some cases the resulting covariance matrix might also be singular, if the

[1] Note that in computer vision literature positive-definite covariance matrices are often denoted as tensors.

considered samples are linearly dependent (*i.e.*, an observed image provides substantial homogeneous regions and edge or 3D information are directly derived from the intensity values) or if the number of samples is less than $D + 1$. In these cases the resulting distribution has no density.

Since covariance matrices can also be singular, they are symmetric and positive semi-definite by definition with $\Sigma \in Sym_D$, where Sym_D denotes the space of these matrices. In this vector space a $D \times D$ covariance matrix can be fully described by $D(D+1)/2$ distinct entries. To ease the working with covariance matrices, we require that the covariance matrices have to be symmetric positive-definite (SPD) matrices with $\Sigma \in Sym_D^+$. In order to handle singular cases, *e.g.*, computing covariance descriptors within homogeneous image regions, we introduce a commonly used regularization[1] with $\Sigma = \Sigma + \epsilon \mathbf{I}_D$, where $\mathbf{I}_D$ is the D-dimensional identity matrix and $\epsilon =$ 1e-6. $\Sigma \in Sym_D^+$ is then a strictly symmetric ($\Sigma = \Sigma^T$) and positive-definite matrix, satisfying $\mathbf{y}^T \Sigma \mathbf{y} > 0$ for all $\mathbf{y} \in \mathbb{R}^D$. These SPD matrices are also often denoted as tensor matrices [Pennec et al., 2006].

3.4.2 Working with Symmetric Positive-Definite Matrices

Since there exists no closed-form solutions for integrating the probability density function $f_X(\mathbf{x})$ (see (3.1)), various methods for approximation have been proposed in the past. The methods range from computationally expensive Monte-Carlo simulation methods, over approximations, such as truncated Taylor series expansion, to prominent techniques, provided by differential geometry [Boothby, 1975]. In the following section we briefly discuss related methods in the context of feasible representations, where well-established distance measurements can then be computed efficiently.

Computation on Riemannian Manifolds

Differential geometry based approaches are widely adopted to compute distance between observed covariance matrices on Riemannian manifolds [Förstner and Moonen, 1999, Goldberger et al., 2003, Arsigny et al., 2007]. In [Pennec et al., 2006, Dryden et al., 2009] detailed summaries are given for derived metrics and methods of covariance interpolation and filtering.

For a SPD matrix $\Sigma \in Sym_D^+$, there exist two fundamental operations, the matrix exponential and the logarithm, which can be easily computed by utilizing the spectral decomposition of the considered matrix. Common Euclidean vector space operations, such as addition and subtraction, are then represented via exponential and logarithm maps on the Riemannian manifold.

[1] Note that even a small amount of noise, added to the considered samples, might largely avoid singular covariance matrices.

Let $\Sigma = L\Lambda L^T$ be an ordinary spectral decomposition, where L is an orthonormal matrix, and Λ is real and diagonal. The exponential series is then given by

$$\exp(\Sigma) = \sum_{k=0}^{\infty} \frac{\Sigma^k}{k!} \quad = \quad L \exp(\Lambda) L^T, \tag{3.2}$$

where $\exp(\Lambda)$ is the diagonal matrix composed of the eigenvalue's exponentials. As in the scalar case, the matrix logarithm is defined as the inverse of the exponential. The logarithm series of the matrix Σ can be thus written as

$$\log(\Sigma) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} (\Sigma - \mathbf{I}_D)^k \quad = \quad L \log(\Lambda) L^T, \tag{3.3}$$

where $\log(\Lambda)$ is the diagonal matrix, that defines the eigenvalue logarithms. Note that the exponential series is defined for any symmetric matrix in Sym_D due to the one-to-one correspondence between the symmetric matrix and the tensor. In contrast, the inverse function, the matrix logarithm, is defined solely for SPD matrices. However, in our case this might be guaranteed by the applied regularization step. Based on the specific properties of the matrix exponential and logarithm, we can now define a vector space structure on tensors, providing a simple vectorization of SPD matrices or metrics providing valid distance computations.

In fact, the space of SPD matrices Sym_D^+ lies on a connected Riemannian manifold $\mathcal{M}$ describing a Lie group. A manifold can be seen as a topological space, that is locally like an isomap to a Euclidean vector space since every point on the manifold has a neighborhood for which there exists an isomorphism, mapping the neighborhood to $\mathbb{R}$ [Pennec et al., 2006]. For a differentiable Riemannian manifold, we can define the derivatives of curves on the manifold. The derivatives at a point $X \in \mathcal{M}$, defined on the manifold, live in a vector space, which describes the tangent space at that point. This tangent space can be thus clearly identified by the space of SPD matrices Sym_D^+. The distance between two points $d(X, Y)$ with $Y \in \mathcal{M}$ is then given by the length of a curve, connecting these points on the manifold. The minimum length is also referred to as the geodesic.

Two invariant Riemannian metrics are typically adopted to compute the geodesic between samples of SPD matrices, namely the Log-Euclidean and the affine-invariant metric.

Log-Euclidean Metric. Under the Log-Euclidean Riemannian metric the Lie group structure of SPD matrices can be extended to a Lie algebra. The Lie algebra also defines a vector space on tensors, that can be then described by a Euclidean metric. The geodesic between two SPD matrices Σ_1 and Σ_2 is given by

$$d_{LE}(\Sigma_1, \Sigma_2) = \| \log(\Sigma_1) - \log(\Sigma_2) \|_F \ . \tag{3.4}$$

From (3.4) it can be seen that the Log-Euclidean metric directly corresponds to Euclidean metrics, such as the Frobenius norm (L^2 norm applied to matrices), in the domain of matrix logarithms, since the shortest distance between the two points in the manifold is given by straight line. As shown in [Pennec et al., 2006] this metric overcomes the problem of computational limitations while maintaining the theoretical properties. In addition, it is widely adopted to the exact estimation of statistics for a number of SPD matrix samples [Pennec et al., 2006, Dryden et al., 2009].

Since $\log(\Sigma)$ provides a vector space structure under the Log-Euclidean metric of a given SPD matrix Σ, we can unfold $\log(\Sigma)$ into a feature vector without loosing any information. Considering a region covariance matrix $\Sigma \in Sym_D^+$, we apply the Log-Euclidean mapping according to (3.3). This concept has been extensively applied to learning incremental subspaces in tracking [Li et al., 2008], detection [Wu and Nevatia, 2008] and even semantic classification [Kluckner et al., 2009b].

Affine-Invariant Metric. The second metric defined on Riemannian manifolds is referred to as the affine-invariant distance [Porikli et al., 2006, Pennec et al., 2006]. Given two SPD matrices Σ_1 and Σ_2, the metric $d_{AI}(\Sigma_1, \Sigma_2)$ computes the length of the geodesic connecting these two points on the manifold with

$$d_{AI}(\Sigma_1, \Sigma_2) = \| \log(\Sigma_1^{-\frac{1}{2}} \Sigma_2 \Sigma_1^{-\frac{1}{2}}) \|_F \ . \tag{3.5}$$

Note that any transformation of the form $\Sigma' = Q\Sigma Q^T$ applied to the covariance matrix Σ, satisfying Sym_D^+, does not affect the computed distance under this metric. As shown by Förstner and Moonen [Förstner and Moonen, 1999] Equation (3.5) can be solved in closed-form based on generalized eigenvalues λ_i that are computed for $\Sigma_1^{-\frac{1}{2}} \Sigma_2 \Sigma_1^{-\frac{1}{2}}$. The metric is defined as

$$d_f(\Sigma_1, \Sigma_2) = \sqrt{\sum_{i=1}^{D} \log^2 \lambda_i(\Sigma_1, \Sigma_2)} \ . \tag{3.6}$$

The definition of this metric guarantees the triangle inequality, symmetry and positivity for SPD matrices [Förstner and Moonen, 1999]. However, compared to the Log-Euclidean metric, the affine-invariant metric involves matrix inversions, square roots, logarithms and exponentials, and has therefore an increased computational complexity.

Kullback-Leibler Divergence. Both the Log-Euclidean and the affine-invariant metric enable a similarity computation between SPD matrices that represent solely the second order moments. In this work we aim to additionally consider the discriminative information given by the mean vector.

The Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951] is a popular method to compute similarity between two fully parametrized normal distributions. Let $f = \mathcal{N}_D(\mu_f, \Sigma_f)$ and $g = \mathcal{N}_D(\mu_g, \Sigma_g)$ be two D-dimensional multivariate normal distributions, the similarity measurement between f and g, satisfying positivity, is defined as

$$d_{KL}(f,g) = \frac{1}{2}\left(\log\frac{|\Sigma_g|}{|\Sigma_f|} + tr\left(\Sigma_g^{-1}\Sigma_f\right) - D + \right. \\ \left. + (\mu_f - \mu_g)^T \Sigma_g^{-1} (\mu_f - \mu_g)\right) . \tag{3.7}$$

Note that KL divergence is not symmetric $d_{KL}(f,g) \neq d_{KL}(g,f)$, however, a symmetric similarity measurement can be obtained by using the sum with $d_{KL}(f,g) + d_{KL}(g,f)$. Since there exists no closed-form solution for computing the similarity between mixtures of Gaussians, Monte Carlo simulation methods are widely adopted to approximate the divergence [Goldberger et al., 2003]. Similar to the KL divergence the Bhattacharyya distance [Bhattacharyya, 1945] can be used to compute the distance between two Gaussian distributions.

Shape of Gaussians. Another elegant way to compute distances between first and second order statistics is described in [Gong et al., 2009], where the authors propose to utilize a feature descriptors based on Shape of Gaussians (SOG). Based on the mean vector μ and a decomposed covariance matrix $\Sigma = AA^T$, representing a multivariate Gaussian distribution, a positive definite lower triangular (PDLT) matrix

$$M = \begin{pmatrix} A^T & \mu \\ 0 & 1 \end{pmatrix}^T \tag{3.8}$$

can be constructed. Let $\mathbf{x}_0$ be a D-dimensional random vector drawn from a Gaussian distribution $\mathbf{z} \sim \mathcal{N}_D(0, \mathbf{I}_D)$ and $\mathbf{x}_1$, $\mathbf{x}_2$ be samples corresponding to the normal distributions $\mathbf{x}_1 \sim \mathcal{N}_D(\mu_1, \Sigma_1)$ and $\mathbf{x}_2 \sim \mathcal{N}_D(\mu_2, \Sigma_2)$. Then, the following mapping functions can be defined:

$$\begin{pmatrix} \mathbf{x}_2 \\ 1 \end{pmatrix} = M_2 \begin{pmatrix} \mathbf{z} \\ 1 \end{pmatrix} = M_2 M_1^{-1} \begin{pmatrix} \mathbf{x}_1 \\ 1 \end{pmatrix} , \tag{3.9}$$

where M_1 and M_2 denote the constructed PDLT matrices for the considered Gaussian distributions. Since PDLT matrices are closed under matrix multiplications and inverse computations, the resulting matrices are again PDLT matrices. Hence, it can be seen that the mapping from $\mathbf{x}_1$ to $\mathbf{x}_2$ with $M_2 M_1^{-1}$ is performed with a PDLT matrix. Considering the affine-invariant metric defined in (3.5) the geodesic length between two PDLT matrices are simply computed with

$$d_{SOG}(\mathcal{N}_D(\mu_1, \Sigma_1), \mathcal{N}_D(\mu_1, \Sigma_1)) = \| \log(M_1^{-1} M_2) \|_F \ . \tag{3.10}$$

Approximation of Normal Distributions

As shown in the last sections, exponential and logarithmic mappings to Riemannian manifolds can be utilized to find meaningful vector spaces, where well-established norms measure a valid distance. Unfortunately, even Riemannian manifolds mappings do not provide a straightforward solution for the combination of first and second order statistics, which is of interest in case of semantic classification where, *e.g.*, the mean vector also provides an important discriminative measurement. We therefore pursue an efficient technique to represent first and second order statistics directly on a Euclidean vector space, where distance measurements can be then applied. Our concept relies on approximating the mean and covariance matrix directly on a Euclidean vector space by utilizing the constructive definition of multivariate Gaussian distributions. While computationally complex Monte Carlo methods are commonly exploited to accurately simulate a given non-linear function with $1/N \sum_i \mathbf{x}_i$, where $\mathbf{x}_i$ are correlated samples and $N \rightarrow \infty$, we are interested in finding a more efficient, however adequate, approximation of a fully parametrized multivariate distribution.

In the context of state prediction, Julier and Uhlmann [Julier and Uhlmann, 1996] proposed the concept of the unscented transformation (UT), where first and second order statistics are propagated through the system using non-linear transformations. Thus, the UT can also be seen as an efficient method for approximating the statistics of a random variable that undergoes a non-linear transformation. The concept of UT is mainly founded on the intuition that it is easier to approximate a normal distribution than it is to approximate an arbitrary non-linear function or transformation. Although UT is similar to Monte Carlo methods, no random sampling is used and only a reduced set of deterministically chosen points are required to approximate the source distribution. In the following we outline the concept of Sigma Points, and show how to construct a meaningful feature representation to integrate various low-feature cues for the task of semantic interpretation.

3.4.3 Generating Correlated Samples

Consider an arbitrary sample vector $\mathbf{x} = [x_1, x_2, \ldots, x_D]^T$, where each scalar element is normally distributed with $x_i \sim \mathcal{N}(\mu_i, \sigma_i)$ and $i = \{1 \ldots D\}$. Since a normal distribution is closed under linear transformations, we can write a linear combination of the random variables under the assumption, that the scalars x_i are independent and identically distributed with

$$\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_D x_D \ \sim \mathcal{N}(\sum_i \alpha_i \mu_i, \sum_i \alpha_i^2 \sigma_i^2). \tag{3.11}$$

More generally, we can rewrite (3.11) for multivariate distributions in vector notation as

$$A\mathbf{x} \ \sim \mathcal{N}_D(A\mu, AA^T), \tag{3.12}$$

where $A \in \mathbb{R}^{D \times D}$ satisfies $AA^T = \Sigma$. Since our computed covariance matrix Σ is SPD we can always decompose the matrix according to $\Sigma = L\Lambda L^T$, where $L \in \mathbb{R}^{D \times D}$ is orthogonal and $\Lambda \in \mathbb{R}^D$ is a positive diagonal matrix.

Recall that for the univariate case the probability density function of a random variable is symmetric around the mean μ and that a multivariate normal distribution can be seen as a linear combination of univariate distributions. We can thus map a multivariate normal random variable $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ to the standard normal $\mathbf{z} = [z_1, z_2, \ldots, z_D]^T$ with $z_i \sim \mathcal{N}(0, 1)$ and $i = \{1 \ldots D\}$, and vice versa according to the constructive definition with

$$\mathbf{z} = \Sigma^{-1/2}(\mathbf{x} - \mu) \quad \Longleftrightarrow \quad \mathbf{x} = \Sigma^{1/2}\mathbf{z} + \mu. \tag{3.13}$$

According to (3.13), a sample vector $\mathbf{x} \sim \mathcal{N}_D(\mu, \Sigma)$ can now be seen as a correlated random vector mainly constructed over a collection of independent and normally distributed variables drawn from $\mathbf{z} \sim \mathcal{N}_D(0, \mathbf{I}_D)$. Note that the SOG concept (3.9) makes use of the constructive definition. Figure 3.3 depicts these relations for the case with $D = 2$.

Moreover, from (3.13) we can see that the non-linear transformation is mainly based on the computation of the matrix square root of Σ. Although the square root of a matrix is not unique, we can use any square root since one root can be mapped to another by applying an orthonormal transformation with $LL^T = \mathbf{I}$, where L is an orthonormal matrix. Thus, we can decompose the SPD matrix Σ according to

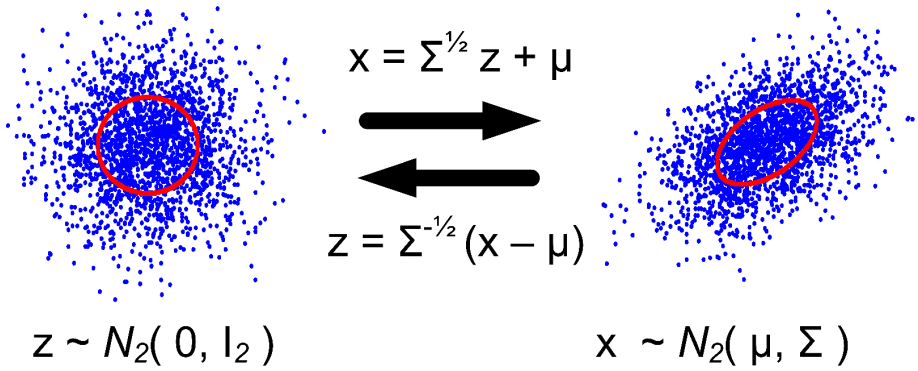


Figure 3.3: Transformation between individual samples, drawn from the standard normal, and the correlated random vector for a 2D case. The blue dots denote arbitrary samples, while the red points are samples defined on the unit circle. The transformation is given for $\mu = [0, 0]^T$ and $\Sigma = [1\ 0.5; 0.5\ 1]$.

$$\begin{aligned}\Sigma &= L\Lambda L^T \\ &= L\Lambda^{1/2}\Lambda^{1/2}L^T \\ &= L\Lambda^{1/2}(\Lambda^{1/2})^T L^T \\ &= L\Lambda^{1/2}(L\Lambda^{1/2})^T \\ &= AA^T. \end{aligned} \tag{3.14}$$

From (3.14) we can observe that $A = L\Lambda^{1/2}$, and therefore derive a formulation that defines random normal vectors to have multivariate normal distribution $\mathbf{x} \sim \mathcal{N}_D(\mu, \Sigma)$ with

$$\mathbf{x} = A\mathbf{z} + \mu = (L\Lambda^{1/2})\mathbf{z} + \mu, \tag{3.15}$$

where $\mathbf{z} \sim \mathcal{N}_D(0, \mathbf{I}_D)$. Furthermore, considering both (3.14) and (3.15), we can notice that the matrix square root computation is not restricted to roots providing an orthogonality constraints (*e.g.*, the singular value decomposition (SVD)) which are computationally more expensive. In fact, a stable method like the Cholesky factorization can be applied efficiently to compute the matrix square root $\Sigma = AA^T$, where A is then a lower triangular matrix.

So far, we have shown that random vectors drawn from an arbitrary normal distributions can be transformed to the standard normal and vice versa (see Figure 3.3), whereas

the transformation clearly reflects the characteristics of the statistical moments. In particular, in this work we are interested in computing a finite set of correlated samples, that allows us to represent the properties of available mean and covariance information.

3.4.4 Sigma Points Representation

Julier and Uhlmann [Julier and Uhlmann, 1996] proposed the unscented transform for the approximation of individual normal distribution by deterministically sampling instead of finding an exact, however, closed-form solution of the non-linear density function in a manifold [Pennec et al., 2006]. As shown in [Julier and Uhlmann, 1996, Julier and Uhlmann, 1997], the UT provides an efficient estimator for the underlying probability distribution and was successfully applied to unscented Kalman filtering, where it overcomes the drawbacks of truncated (second order) Taylor expansions used for extended Kalman filtering. In the D-dimensional case the UT relies on constructing a small set of $2D + 1$ specific vectors $\mathbf{s}_i \in \mathbb{R}^D$, also referred to as Sigma Points [Julier and Uhlmann, 1996]. Considering the constructive definition, given in (3.15), the generation for the set of Sigma Points is defined as follows:

$$\mathbf{s}_0 = \mu \qquad \mathbf{s}_i = \mu + \alpha(\Sigma^{1/2})_i \qquad \mathbf{s}_{i+D} = \mu - \alpha(\Sigma^{1/2})_i, \tag{3.16}$$

where $i = \{1 \dots D\}$ and $(\Sigma^{1/2})_i$ defines the i-th column of the required matrix square root $\Sigma^{1/2}$. The scalar α defines a constant weighting for the elements in the covariance matrix and is set to $\alpha = \sqrt{2}$ for Gaussian input signals [Julier and Uhlmann, 1996]. In contrast to Monte Carlo methods, where test vectors are taken randomly from a standard normal, the construction of the Sigma Points can be seen as an efficient mapping of a specified set of test vectors $\mathbf{t}_i \in \mathbb{R}^D$ that deterministically sample the intersections of an unit hypersphere with a D-dimensional Cartesian coordinate system. Due to symmetry and positive definiteness of the regularized covariance matrices, the efficient and stable Cholesky factorization can be applied to compute the matrix square root of Σ. In principle any method for square root factorization could be used, however, the Cholesky decomposition requires the least mathematical operations with complexity $\mathcal{O}(n^3/3)$.

Figure 3.4 illustrates the specified sampling of the test vectors and the mapping for a simplified 2D case. Here, the mean vector $\mathbf{t}_0 = \mu$ represents the origin in a Euclidean vector space. The resulting feature representation $\mathcal{S} = (\mathbf{s}_0, \mathbf{s}_1, \dots, \mathbf{s}_{2D})$ is obtained by concatenation of the individual Sigma Points and captures both, first and second order statistics. Each of these generated vectors $\mathbf{s}_i \in \mathbb{R}^D$ describes a Euclidean space, allowing element-wise distance computations between corresponding samples of a given distribution. The Sigma Points construction is summarized in Algorithm 3.1. From this algorithm

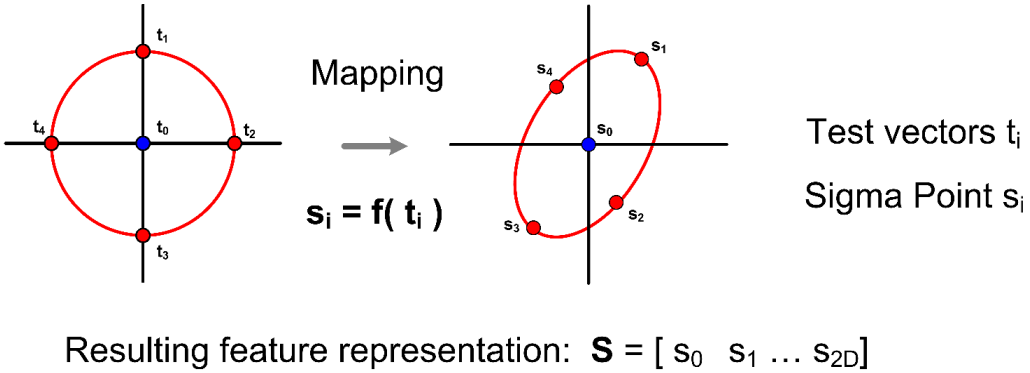


Figure 3.4: Sigma Points generation for a 2D case ($D = 2$): The set of Sigma Points $s_{0,\ldots,2D}$ reflects the characteristic of the non-linear function $f(\cdot)$. The test vectors $\mathbf{t}_{1\ldots2D}$ are the canonical bases given by the dimension D. The center point $\mathbf{s}_0$ directly corresponds to the $\mathbf{t}_0$ and the mean vector μ, respectively, which can be seen as an offset, defining a translation in Euclidean space.

one can observe that the construction solely consists of basic vector and matrix operations making an implementation very simple. In [Julier and Uhlmann, 1996] the authors proved that the computed statistics for the points $\mathbf{s}_i$ accurately capture the original information about μ and Σ up to third order for Gaussian and up to second order for non-Gaussian inputs.

Algorithm 3.1 Construction of the Sigma Points.

Require: Mean vector μ and covariance matrix Σ

1: Perform regularization and weighting $\Sigma = \alpha\Sigma + \epsilon\mathbf{I}$
2: Use Cholesky decomposition to compute matrix square root $\Sigma = AA^T$
3: Compute $\mathbf{s}_i$ according to (3.16)
4: Construct the set of Sigma Points $\mathcal{S} = (\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_{2d})$

3.4.5 Feature Representation for the Semantic Interpretation

In our case we extract the statistics directly from the image content, incorporating arbitrary low-level feature cues. Let $\mathcal{I} \subseteq \mathbb{R}^{W\times H\times D'}$ be an one- or three-dimensional intensity image, and let $F \subseteq \mathbb{R}^{W\times H\times D}$ be a feature image generated from $\mathcal{I}$ with

$$F(\mathbf{p}) = \Phi(\mathcal{I}(\mathbf{p})), \tag{3.17}$$

where $\mathbf{p} = (p_x, p_y)$ is a pixel position defined the image $\mathcal{I}$. In our case, the feature generator function $\Phi(\cdot)$ provides mappings, such as color space conversions, derivative and

height field computations. Then, the feature vector $F(\mathbf{p}) \in \mathbb{R}^D$ corresponds to vector consisting of extracted attributes at the image coordinate (p_x, p_y) in F. These vectors are not restricted to normalized value ranges and include usually appearance, basic texture descriptions, but even spatial attributes, magnitudes or angles can be used for region description [Porikli et al., 2006].

Given the feature image F, any extracted covariance descriptor of an arbitrary region $R \subset F$ results in a covariance matrix $\Sigma \in Sym_D$:

$$\Sigma = \frac{1}{|R| - 1} \sum_{\mathbf{p} \in R} (F(\mathbf{p}) - \mu)(F(\mathbf{p}) - \mu)^T, \tag{3.18}$$

where $|R|$ denotes the cardinality of considered samples in the region R and $\mu \in \mathbb{R}^D$ is the sample mean vector given by

$$\mu = \frac{1}{|R|} \sum_{\mathbf{p} \in R} F(\mathbf{p}). \tag{3.19}$$

A restriction to non-spatial attributes preserves the scale and rotation invariance because Σ does not capture the ordering of the incorporated attribute vector in the image grid. Applying the Sigma Points construction (see Algorithm 3.1) to μ and Σ yields a feature vector that compactly reflects the characteristics of first and second order statistics[1].

In the following we briefly outline the rapid construction of the statistics using an extended set of integral images. The feature representation clearly can be computed for rectangular regions as well as for areas provided by unsupervised segmentation methods.

3.4.6 Integral Image Structures

Integral image structures, also referred to as summed area tables, are simple, but powerful representations for efficient computation of feature value sums within rectangular regions. By exploiting the integral structure, any region sum can be computed in constant time independently of the region dimension. Viola and Jones introduced integral images over intensity values for rapid object detection [Viola and Jones, 2004], where each pixel of the integral images is the sum of all pixels within a rectangle, defined by the image origin and an investigated pixel. The integral image $\mathrm{Int}(r, c)$ at a coordinate (r, c) is defined as F with

$$\mathrm{Int}(r, c) = \sum_{p_x \leq r} \sum_{p_y \leq c} F(p_x, p_y) \tag{3.20}$$

[1] We have also applied the Sigma Points, extracted from both a parametrized normal distribution [Kluckner et al., 2009a] and generalized structure tensors [Donoser et al., 2010], to the task of visual object tracking, where we used well-established norms to compute for similarity between extracted representations.

for an input image. As shown in various approaches, integral structures can be extended to handle multi-dimensional data. In our approach we exploit integral structures to compute first and second order statistics of rectangular image regions [Tuzel et al., 2006]. Given the feature image F, we first construct D integral images, containing the summed feature values, with

$$S_i(r,c) = \sum_{p_x \leq r} \sum_{p_y \leq c} F(p_x, p_y, i),\ i = \{1 \dots D\}\ . \tag{3.21}$$

In order to compute the covariance matrices we additionally construct integral images providing us with the permuted pairs of tensors of the feature channels according to

$$T_{ij}(r,c) = \sum_{p_x \leq r} \sum_{p_y \leq c} F(p_x, p_y, i)\ F(p_x, p_y, j),\ i,j = \{1 \dots D\}\ . \tag{3.22}$$

Since covariance matrices are symmetric, a reduced set of $D(D+1)/2$ tensor integral images are sufficient for a full construction of the second order moment.

Let $R(p_{x1}, p_{y1}, p_{x2}, p_{y2})$ be a region of interest with a left upper corner (p_{x1}, p_{y1}) and a lower right corner (p_{x2}, p_{y2}), the sums and the sums for the tensors, composed over the feature attributes can be computed with

$$S_i^R = S_i(p_{x1}, p_{y1}) + S_i(p_{x2}, p_{y2}) - S_i(p_{x1}, p_{y2}) - S_i(p_{x2}, p_{y1}),\ i = \{1 \dots D\}\ , \tag{3.23}$$

and

$$T_{ij}^R = T_{ij}(p_{x1}, p_{y1}) + T_{ij}(p_{x2}, p_{y2}) - T_{ij}(p_{x1}, p_{y2}) - T_{ij}(p_{x2}, p_{y1}),\ i,j = \{1 \dots D\}\ , \tag{3.24}$$

respectively. Given the sums S_i^R and sums of tensors T_{ij}^R for a regions R, we use an unbiased estimate of the population variance to compute the covariance matrix of the region R with

$$\Sigma(i,j) = \frac{1}{|R|-1}\left(T_{ij}^R - \frac{1}{|R|} S_i^R S_j^R\right)\ i,j = \{1 \dots D\}\ , \tag{3.25}$$

where $|R|$ again denotes the number of covered pixels within the rectangular region R. The first order statistics of a region R is defined as

$$\mu(i) = \frac{1}{|R|} S_i^R \tag{3.26}$$

with $i = \{1 \dots D\}$. This computation enables a construction of sample mean and covariance matrix with constant time consumptions. The complexity for the integral image

computation is $\mathcal{O}(D^2WH)$. Note that covariance descriptors can also be computed in one pass for arbitrary region shapes, *e.g.*, provided by an unsupervised segmentation, by using additionally computed tensors and lists of coordinates.

3.5 Randomized Forest Classifier

The structure of the proposed Sigma Points representation $\mathcal{S}$ perfectly fits the concept of RF classifiers, where the learning and evaluation strategy is mainly based on comparing randomly selected attributes of an observed feature instances. In the case of semantic interpretation, where a meaningful reference model is missing or at least costly to estimate, similarity measurements, such as the KL-divergence [Goldberger et al., 2003], affine-invariant [Förstner and Moonen, 1999] or Log-Euclidean [Pennec et al., 2006] metric are intractable to use in decision trees directly. We thus outline how the Sigma Points can be applied straightforwardly to a RF framework.

Randomized forests[1], proposed by [Amit et al., 1996, Breiman, 2001], have been successfully applied to various computer vision problems, including clustering [Moosmann et al., 2006], visual tracking by keypoint recognition [Lepetit and Fua, 2006], object categorization [Bosch et al., 2007] or detection [Gall and Lempitsky, 2009], semantic classification [Shotton et al., 2008, Schroff et al., 2008, Brostow et al., 2008] and regression [Criminisi et al., 2010]. While regression forests provide an efficient mapping of high-dimensional data to model parameters, a RF classifier projects real-valued feature instances to probabilistic decisions.

Efficiency in both training and testing, inherent multi-class capability and accuracy make RFs highly applicable to challenging practical problems. These classifiers additionally reduce the problem of overfitting [Amit et al., 1996] and provide robustness to label noise. The handling of label noise is particularly important since massive human interaction is used to generate labeled training maps and human are known to make (minor) mistakes. Moreover, since the trees in the forest are trained and evaluated independently, RFs can be accelerated considerably by using multi-core systems [Sharp, 2008]. This makes them interesting for large-scale tasks or real-time problems.

An RF classifier consists of an ensemble of T binary decision trees , where the nodes of each tree include split criteria that give the direction of branching left and right down the tree until a leaf node in a predefined depth Z is reached. A leaf node $l_i \in L$ then contains the class distribution $P(\mathbf{c}|l_i)$ generated by the target labels of the visible training examples. An averaging over all tree decisions in the forest yields the resulting accumulated class distribution obtained for each evaluated feature instance according to

[1] Randomized forests are also referred to as randomized tree, random forests or random decision trees.

$$P\left(\mathbf{c}|L\right) = \frac{1}{T}\sum_{i=1}^{T} P\left(\mathbf{c}|l_i\right). \tag{3.27}$$

A mode estimation with $\mathbf{c}^* = \arg\max_c P\left(\mathbf{c}|L\right)$ computes the final object class $\mathbf{c}^*$.

A random tree construction proceeds from the root node top-down by splitting the available training subset at each node into tiled left and right feature sets. To rapidly grow each tree of the forest, the split node criteria are learned using only a subset $\mathcal{S}'$ of the whole training data $\mathcal{S}$. For the process of training each feature vector $\mathcal{S}^k \in \mathcal{S}$ is assigned a class label c_k that is provided by the labeled ground truth maps. The learned split criteria in the nodes then maximize the sample-weighted information gain ratio ΔH of the class distribution in the currently available subsets of the training data:

$$\Delta H = -\frac{|\mathcal{S}_r|}{|\mathcal{S}_l + \mathcal{S}_r|} H(\mathcal{S}_r) - \frac{|\mathcal{S}_l|}{|\mathcal{S}_l + \mathcal{S}_r|} H(\mathcal{S}_l)\ . \tag{3.28}$$

Here, $H(\cdot)$ describes the computed Shannon entropy considering the class distribution with $H(\cdot) = -\sum_{i=1}^{|c|} p_i \log p_i$. The terms $|S_r|$ and $|S_l|$ define the cardinality of the feature vectors split to right or left branch. Typically, a greedy strategy is used to optimize the information gain by simply comparing raw channel values [Shotton et al., 2008], linear combinations of two or more feature values [Bosch et al., 2007], or by using sophisticated or task-specific node test [Schroff et al., 2008, Gall and Lempitsky, 2009].

Our classifier implementation follows the learning strategy as proposed by [Bosch et al., 2007], where linear combinations are computed for the entire feature vector in order to estimate the split decision. We have implemented an accelerated technique by simply taking into account the corresponding dimension $\{1 \leq a \leq D\}$ and by randomly selecting two weighted elements $\{1 \leq i, j \leq 2D+1, \forall i, j\ i \neq j\}$ according to

$$\alpha \mathbf{s}_i(a) + \beta \mathbf{s}_j(a) = \begin{cases} > \gamma, & split\ left \\ \leq \gamma, & split\ right \end{cases}. \tag{3.29}$$

The scalar parameters α, β, and γ denote greedy-optimized values that maximize the information gain with respect to the training labels [Shotton et al., 2008]. The optimized split decision, defined in (3.29), can then be seen as a discrimination of projected Sigma Points $\mathbf{s}_i \in \mathbb{R}^D$ in an one-dimensional space. In [Kluckner and Bischof, 2009] we have estimated the split criteria by directly integrating the constructive definition (3.15) into the decision nodes. Compared to the incomplete representation obtained with the Sigma Points, this concept has considerably suffered from overfitting.

After forest construction by only using a subset of the training samples $\mathcal{S}' \subset \mathcal{S}$, each tree is refined with the entire set of feature vectors in order to generate the final leaf node

class distributions. This technique permits a sophisticated handling of a large amount of data and significantly improves the generalization capability [Shotton et al., 2008].

At testing time, we evaluate the classifier at each pixel location by passing down the extracted feature representation in the forest and accumulating the class distributions to obtain $P(\mathbf{c}|L)$. It is evident that other classifiers, such as multi-class boosting [Shotton et al., 2007], multinomial logit modeling [Ranganathan, 2009], etc., could be used instead. For instance, in [Nguyen et al., 2010] we have shown that the Sigma Points representation can be trained and evaluated with linear SVMs.

While learning methods, like SVMs, boosting or logistic regression, are based on computing linear combinations, the structure of the RF classifier enables a very fast testing procedure since the number of evaluations is directly related to the defined tree depth Z. Ordinarily, we construct a forest with $T = 10$ trees, each with a depth of $Z = 15$, hence, an observed feature instance is then evaluated with a low number of 10×15 binary decisions during testing.

3.6 Introducing Segmentation for Object Delineation

So far, we have considered the semantic interpretation at the pixel level (by incorporating small rectangular image regions) due to efficient construction using integral images. While well-defined objects like cars, pedestrians or faces can be successfully detected with bounding boxes, *e.g.*, with [Viola and Jones, 2004] it is obvious that real-world images include many objects in varying scales that cannot be captured with a fixed-scale rectangular prior. In addition, running a sliding window strategy does not incorporate potential relationships between evaluated locations implicitly. Figure 3.5 shows a small part of a classification result for *Graz*. One can notice that the obtained result shows a high granularity with respect to the dominant classes mainly caused by minor classification errors or the lack of missing context information, such as derived object boundaries.

As a next step efficient unsupervised segmentation methods are thus investigated to refine the initial classification results. Unsupervised image segmentation can be generally seen as the task of partitioning a given image into distinct regions. These regions commonly consist of pixels that have a similar characteristic. Recent approaches combine the results of different segmentation methods (by varying the parameter setting of, *e.g.*, the popular mean-shift algorithm [Comaniciu and Meer, 2002] or the graph-based segmentation method by Felzenszwalb and Huttenlocher [Felzenszwalb and Huttenlocher, 2004b]) in order to determine semantically coherent segments of an observed image. These coherent image segments are also denoted as super-pixels [Ren and Malik, 2003].

Recently, Fulkerson *et al.* [Fulkerson et al., 2009] used Quickshift [Vedaldi and Soatto, 2008, Fulkerson and Soatto, 2010] to compute super-pixels for the description of the im-

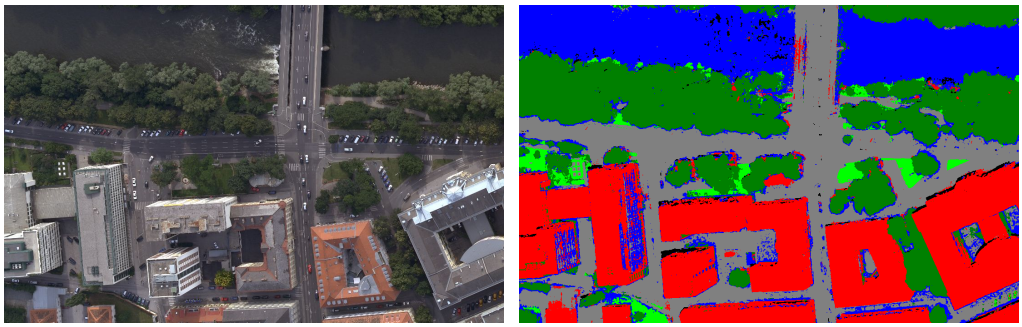


Figure 3.5: An intermediate interpretation result of the proposed pixel-wise classification. The result shows a high granularity with respect to the computed class modes. We thus use an unsupervised segmentation step to obtain an improved semantic labeling.

age content. They showed that these super-pixels accurately preserve object boundaries of natural as well as man-made objects and that subsequent stages benefit from the reduced image grid. Hence, a single super-pixel segmentation offers two important advantages: (i) computed super-pixels can be seen as the smallest units in the image space. All subsequent processing steps can be performed on a reduced graph instead of incorporating the full pixel image grid. The resulting relationships between neighboring super-pixels are stored in an adjacency graph that can be processed efficiently. (ii) Super-pixels are considered as homogeneous regions providing important spatial support: Due to edge preserving capability, each super-pixel describes a coherent part of only one object class. Aggregating data, such as the class distributions provided by the classifier, over the pixels defining a super-pixel compensates for wrong detections and boundary effects.

In Figure 3.6 computed Quickshift super-pixels are given for a small scene of *Graz*. One can notice that a single image partition captures nearly all relevant object boundaries (*i.e.*, transitions between buildings and trees). Additionally, one can see that the super-pixels significantly differ from the rectangular prior which is used for to compute the initial semantic image interpretation.

In order to accurately capture object boundaries within the semantic classification we highlight two simple concepts that make use of the super-pixels.
The first concept relies on a direct computation of the Sigma Points representation for each super-pixel. It involves a conservative super-pixel generation for the training process, but feature extraction and classification can be performed on a reduced image grid.
The second method is based on extracting Sigma Points for small rectangular patches located around each pixel. Contrary to first concept, where the trained classifier directly yields a class distribution for an assembly of connected pixels, a super-pixel is rather treated as a region providing important spatial support. At runtime, the classifier com-

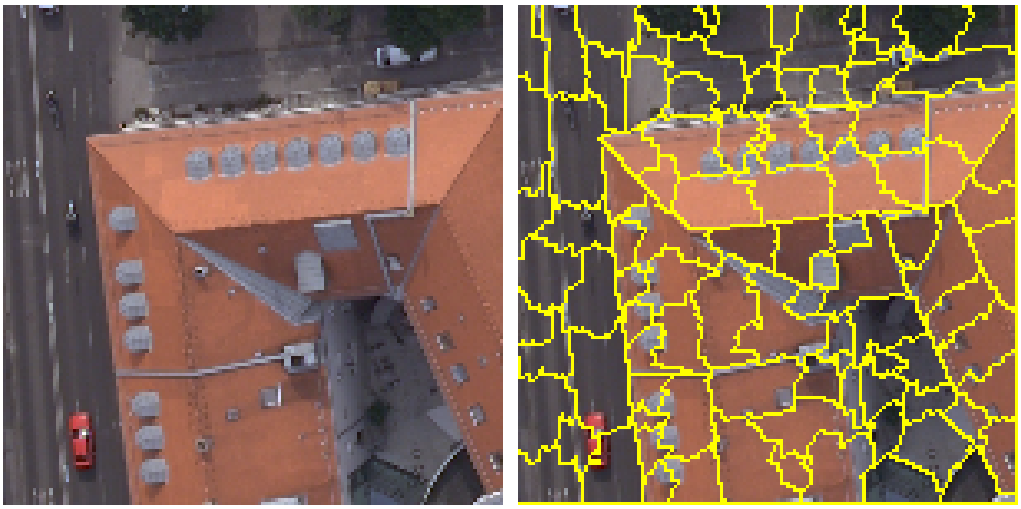


Figure 3.6: Super-pixel segmentation of a small image. Even a single image segmentation ($\sigma = 2.0$, $\tau = 10.0$) preserves important object boundaries, like the transitions between buildings and trees.

putes confidences for each pixel considering a small rectangular neighborhood by using a sliding window approach [Viola and Jones, 2004]. Then, an aggregation of confidences over the super-pixel region yields an averaged class distribution for each segment. This method does not require any segmentation during training (which significantly simplifies the training process) and benefits from a rapid feature computation within regular patches by using powerful integral images.

Spatial Support provided by Multiple Image Partitions. For the sake of completeness, in [Kluckner et al., 2009b] we have shown that the second method, where super-pixels provide spatial support for the confidence aggregation, can be applied to integrate multiple image segmentations. We employed two different segmentation approaches (*i.e.*, the graph-based method proposed by Felzenszwalb and Huttenlocher [Felzenszwalb and Huttenlocher, 2004b] and the mean-shift approach [Comaniciu and Meer, 2002]), which are selected due to public availability, efficiency, popularity and the use of different techniques for image partitioning. A huge pool of probable connected pixels are produced using these methods with varying parameter settings in a first step. For each segmented region, we group the individual pixel-wise classifications yielding a final class distribution for every region. In order to select consistently classified regions, we compute the Shannon entropy over the aggregated class distribution. Taking into account the minimum entropy over all segmentations for each pixel, a final image partition with corresponding class distributions is obtained by assigning the index of the corresponding region. Due to

high computational complexity, the concept cannot be applied to aerial image interpretation. We have thus applied this approach to benchmark images only. For details we refer to [Kluckner et al., 2009b].

3.7 Refined Labeling

Since our local semantic interpretation strategy yields class distributions for each pixel or super-pixel in the observed image independently, we apply a commonly used refinement step to achieve a smooth and spatially consistent class labeling over the entire image space. The problem of obtaining a final consistent labeling of the image scene can be seen as a task of multi-class image segmentation, where each considered pixel value is selected from a predefined pool of class labels. In general the segmentation problem into multiple object classes for an image domain $\Omega \subset \mathbb{R}^2$ can be defined as a minimization problem of the Potts model [Potts, 1952] with

$$\min_{\omega_i, l_i} \left\{ \lambda \sum_{i=0}^{N} \mathrm{Per}(\omega_i; \Omega) + \sum_{i=0}^{N} \int_{\omega_i} |g - l_i|^2 \, dx \right\}, \tag{3.30}$$

where l_i is a probable assignment to the class i and $\bigcup_{i=0}^{N} \omega_i = \Omega$ forms a valid image partition into N classes with $\omega_i \cap \omega_j = 0, \forall i \neq j$. The first term of the functional incorporates the boundary length of the segment ω_i for the regularization, while the second term considers the data at each point x in the segment ω_i. The scalar value λ defines the trade-off between data term and regularization.

In the discrete setting, the segmentation problem into multiple object classes can be written as a minimization problem according to

$$\min_{\mathbf{c}} \left\{ \sum_{p} D(c_p) + \lambda \sum_{p,q} V(c_p, c_q) \right\}, \tag{3.31}$$

where $D(c_p)$ denotes the data term and the pairwise interaction potential $V(c_p, c_q)$. The goal is to infer a final labeling $\mathbf{c}$ that assigns each defined image region $p \in \Omega$ (this could be a single pixel or even a connected part of an image) a label $c_p \in \mathbf{c}$, where the labeling is then both piecewise smooth and consistent with the observed data. While the observed data commonly consists of initially computed parameters, like as probable assignments to specified object classes, the pairwise potentials often describe image information or learned class constellations within an image [Rabinovich et al., 2006].

Given the learned parameters for the CRF formulation (*i.e.*, unary and pairwise potentials), we seek for an optimized final labeling $\mathbf{c}^*$ that minimize the energy defined

in (3.32). This is also referred to as inference. Although the problem of multi-class image segmentation is NP-hard (a two class problem can be solved exactly), there exists a variety of algorithms to compute a solution approximately. In the context of discrete optimization Boykov *et al.* [Boykov et al., 2001] proposed the concept of α-expansion to simplify the problem to a set of binary labeling problems. In [Komodakis and Tziritas, 2007], the authors used linear programming to find an approximate labeling via graph cuts.

Inference is typically performed on a four-connected neighborhood image graph, which seems suitable for most practical applications with respect to computational complexity [Verbeek and Triggs, 2007, Shotton et al., 2007, Rabinovich et al., 2006, Pantofaru et al., 2008]. Kohli *et al.* [Kohli et al., 2008] proposed to directly consider multiple image partitions within the optimization for the task of semantic image segmentation. Similar to the labeling proposed in [Fulkerson et al., 2009], we exploit the adjacency graph, extracted from a single super-pixel segmentation, to enforce an evident final class labeling with low computational costs.

Let $G(\Omega, E)$ be an adjacency graph with a super-pixel node $\omega_i \in \Omega$ and a pair $(\omega_i, \omega_j) \in E$ be an edge between the segments ω_i and ω_j, then an energy can be formulated with respect to the class labels $c_i \in \mathbf{c}$. Considering an adjacency graph the minimization problem can be defined as

$$\min_{\mathbf{c}} \left\{ \sum_{\omega_i \in \Omega} D(\omega_i|c_i) + \lambda \sum_{(\omega_i,\omega_j)\in E} V(\omega_i, \omega_j|c_i, c_j) \right\}, \tag{3.32}$$

where $D(\omega_i|c_i)$ expresses the unary potential of a super-pixel node. In case of classification refinement, $\mathbf{c}$ represents a labeling of the adjacency graph that assigns each graph node ω_i a label c_i. The unary potential $D(\omega_i|c_i) = -\log(P(\omega_i))$ denotes the class distribution obtained for a super-pixel ω_i, either by aggregating confidences from the pixels or by directly estimated with the RF classifier. The scalar λ again controls the influence of the regularization. In order to consider the region sizes in the minimization process (favoring larger regions), we compute the pairwise edge term $V(\omega_i, \omega_j|c_i, c_j)$ between the both super-pixels ω_i and ω_j by taking into account the number of common boundary pixels and estimated mean color distances. The pairwise potentials for ω_i and ω_j are computed according to

$$V(\omega_i, \omega_j|c_i, c_j) = \frac{b(\omega_i, \omega_j)}{1 + g(\omega_i, \omega_j)} \delta(c_i \neq c_j), \tag{3.33}$$

where $b\,(\omega_k, \omega_j)$ computes the number of common boundary pixels of the super-pixels, $g(\omega_k, \omega_j)$ defines the L^2 norm of the mean color distance vector and $\delta(\cdot)$ is the zero-one indicator function that encodes label jumps between c_i and c_j.

Inference. In this work the initial class labeling is optimized by applying α-expansion moves [Boykov et al., 2001]. Although the primal-dual optimization [Komodakis and Tziritas, 2007] obtains identical results much faster this concept suffers from enormous memory requirements. Note that the energy minimization problem can be easily extended to the corresponding graph, where each pixel denotes a node with four or eight direct neighbors. The size of the resulting graph and thus the computational complexity drastically increases, with number of involved neighboring nodes.

3.8 Experiments on Benchmark Datasets

A first experimental setup demonstrates the proposed semantic interpretation applied to standard benchmark datasets.

Due to the compact representation we exploit several low-level feature cues to describe small spatial neighborhoods. In our case a single low-level feature cue might be one of the color channels, derived filter responses or computed elevation measurements in case of aerial image interpretation. According to the applied feature cues we construct the required integral images for the computation of the covariance matrix descriptors. Having both the mean vector and covariance matrix permits the construction of the Sigma Points feature instances by following the algorithm described in Section 3.4.4. These instances are then used for classifier construction and evaluation, where the RF classifier maps extracted feature vectors to class distributions. The final refinement step takes the probable class assignments as input and provides a full semantic interpretation of the image content.

We evaluate our proposed semantic classification scheme on common benchmark datasets, like the MSRC [Winn et al., 2005, Shotton et al., 2007] and the VOC2007 [Everingham et al., 2007] image collections, but also on the eTRIMS image database [Korč and Förstner, 2009] to demonstrate a broad applicability. Each dataset contains many object classes representing *building*, *grass*, *tree*, *cow*, *sheep*, *water*, *book*, etc. with a large intra-class variability and challenging changes in scale and illumination. Since our approach works in a supervised manner, labeled ground truth data is required to construct and evaluated the classifiers. We thus extract the target classes directly from the pixel-wisely labeled ground truth maps that are offered with the datasets.

In contrast to the MSRC images the ground truth annotations of the VOC2007 and eTRIMS images are provided with a higher degree of accuracy and also include a void class, which encodes border regions between object classes, ambiguous image areas or objects that are not in the pool of labeled object classes. Note that the void-labeled pixels are ignored during training and testing procedures. Figure 3.7 shows some MSRC images and the ground truth labeling. The following paragraphs outline the characteristics of the

Figure 3.7: Some images and ground truth labels of the MSRC dataset: The first row shows typical sample images. The second row depicts the corresponding labeling used for training and testing. The void class is represented with black pixels is not considered in our experiments.

benchmark datasets.

MSRCv1

The MSRCv1 dataset [Winn et al., 2005] consists of 240 images with 9 on the pixel level labeled object classes. For the training and the testing procedure we randomly split the dataset into 120 training and 120 test images. Due to randomness of this splitting procedure, we independently repeat the experiments 20 times to obtain meaningful averaged classification rates. An evaluation on this dataset provides results for a comparison to state-of-the-art approaches [Schroff et al., 2008, Verbeek and Triggs, 2007].

MSRCv2

The MSRCv2 dataset is an extension of the MSRCv1 with an increased number of images and object classes. The annotations include 21 object classes. The experiments on this dataset are performed following the train/test splits as suggested in [Shotton et al., 2007]. A number of 276 images are used for training and the remaining 256 for testing.

VOC2007

The VOC2007 dataset [Everingham et al., 2007] provides extremely challenging images with 20 labeled object classes and a background class. Compared to the MSRC collec-

tions, the dataset includes more images at higher resolution. For our experiment we use 422 images for training and the remaining 210 for testing as suggested in the train/test splitting files [Everingham et al., 2007].

eTrims

The eTrims dataset [Korč and Förstner, 2009] includes 60 challenging terrestrial images showing various building facades that are embedded in different urban environments. We use the labeled ground truth data for the object classes (*building*, *ground*, *sky*, *vegetation*) and repeat the interpretation for different random sets of training images in order to determine a reliable classification accuracy.

3.8.1 Evaluation Metric

We evaluate the classification accuracy with respect to the *PASCAL* VOC2007 segmentation taster challenge protocol [Everingham et al., 2007]. For each pixel in the test images, we predict the class of the object containing that pixel. The provided ground truth labels are then used to evaluate the performance. The classification accuracy is computed across the number of object classes. For each class separately, we compute the score that gives the correct classified percentage of pixels. More precisely, for each object class the number of correctly classified pixels is divided by the number of ground truth pixels. We additionally compute the percentage of correctly pixels over all classes. In order to complete the performance evaluation we give numbers for computation times[1], that are recorded for different processing steps.

3.8.2 Sigma Points and Randomized Forest

In this initial experimental setup we investigate the performance of using Sigma Points within the powerful RF classifiers. As a first experiment different types and combinations of feature channels are integrated by using the proposed Sigma Points representation. Collected feature instances are then trained and evaluated with the RFs. In a second experiment we vary the forest size in order to find a reliable trade-off between accuracy and evaluation time. It is obvious that a short evaluation time is essential for processing a large amount of data. For the initial experiments we use the MSRCv2 dataset with 21, at the pixel level labeled, classes to provide results for a comparison to state-of-the-art approaches.

[1] We use a standard hardware: Dual-Core Processor with 3 GBytes of memory, GeForce GTX 9800, C++ Implementation compiled on Windows XP/Visual Studio 8.0

In order to capture an improved invariance to shape deformations and to synthetically increase the amount of training samples we apply small artificial affine distortions to the training images. The transformed images are then used to extract the Sigma Points instances. Each tree in the RF is initially constructed with a set of 60000 randomly selected feature vectors. During the training phase, each split decision in the nodes of the tree is estimated by maximizing the information gain over 1000 greedy iterations. To obtain precise split decisions near the leaf nodes, we additionally weight the number of performed iterations according to the corresponding node's depth in the tree. After the initial tree construction, the class probabilities in the leaf nodes are refined by passing down the entire set of extracted feature instances. Note that this set may consists of several millions of extracted instances. Due to unbalanced labeling of the training data (the number of extracted representations significantly varies for the object classes), we apply an inverse weighting by taking into account the number of extracted instances for each class to simulate a balanced dataset [Shotton et al., 2008]. At evaluation time we use a symmetric content-padding for all test images in order to obtain valid class assignment near the image borders.

Combination of Features

As commonly suggested in local or patch-based classification approaches [Shotton et al., 2007] we extract the Sigma Points feature representation in a spatial neighborhood of 21×21 pixels. The training data is sparsely collected on a 5×5 image grid and the corresponding target labels are extracted by considering the available ground truth images.

Due the absence of 3D information in this experiment we use different combinations of color spaces[1] and filter responses (we use Sobel filter masks). For each feature combination we extract approximately 1.5 million feature instances in a first step. Due to randomness of the classifier, we repeat the training and evaluation procedure 20 times in order to obtain meaningful performance measurements for each parameter setup. The highest classification accuracy (we consider the averaged class rate) is obtained for an integration of RGB-color and the absolute values for first and second order derivatives computed in both horizontal and vertical direction: $F(x,y) = [R \;\; G \;\; B \;\; d_x \;\; d_y \;\; d_{xx} \;\; d_{yy} \;\;]^T$. Note that the Sigma Points feature vectors have a dimension of $D(2D+1)$, if D low-level feature cues are integrated. In the specific case we thus obtain a feature vector with 105 attributes that represent the characteristics of the mean and the covariance matrix.

A summary of the obtained classification accuracies and evaluation times for a forest size of $T = 10$ and $Z = 15$ is given in Table 3.8. We additionally compare the rates to reported intermediate evaluation results of several state-of-the-art approaches. It can be

[1] We use the OpenCV color space converter, which uses a normalization scheme to $\{0 \dots 255\}$.

R G B	L a b	d_x,d_y	d_{xx},d_{yy}	Pixel Avg [%]	Class Avg [%]	Eval Time [s]
		X		35.17	28.53	0.35
X				50.96	39.90	0.32
	X			48.08	36.90	0.34
X		X		56.67	45.64	0.41
X		X	X	**59.56**	**45.99**	0.49
[Schroff et al., 2008], only color				54.00	n.a.	n.a.
[Schroff et al., 2008], color+HOG				69.90	n.a.	n.a.
[Lazebnik and Raginsky, 2009], initial model				53.26	40.65	n.a.
[Shotton et al., 2008]				14.80	24.10	n.a.

Table 3.1: A quantitative evaluation of the interpretation workflow by using different combinations of integrated low-level feature cues. Note that the given values are computed for the raw classification step without using a refinement. It can be clearly seen that RGB color in combination with first and second order derivatives obtains the best results. In addition, a comparison to reported intermediate results of state-of-the-art approaches demonstrates a similar performance.

clearly seen that our direct approach is able to keep up with the compared methods. A combination of RGB and first/second order derivatives yield the best performance with an averaged evaluation time of about 0.5 seconds per image. Figure 3.8 depicts the averaged class accuracies as a function of varying tree depths. From Figure 3.8 we can notice that tree depths of 15 to 20 obtain a reliable classification result for different combinations of low-level features. Moreover, the additional integration of second-order derivatives only slightly improves the accuracy of the overall result considering the evaluation on 21 object classes.

Interpretation of Terrestrial Images. Mainly inspired by the recent work of Drauschke and Mayer [Drauschke and Mayer, 2010], we apply the proposed classification to the task of initial terrestrial facade interpretation. Similar to the MSRC and VOC images, we extract the Sigma Points feature vectors at sampled image locations by considering a spatial neighborhood of 21×21 pixels. The target labels for the 4 object classes are extracted from the ground truth images. For each run of the experiment, a total number of approximately $400K$ feature instances is extracted for the training process. These instances are trained with a forest, consisting of $T = 10$ trees, each with a maximum depth of $Z = 15$. Due to a missing train/test splitting setup, we repeat the feature extraction, the training and the testing procedure 20 times in order to obtain a meaningful evaluation. The ob-

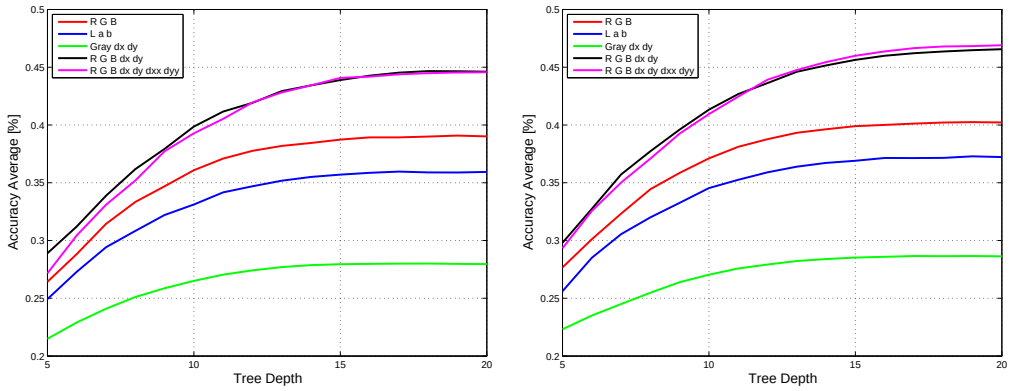


Figure 3.8: Obtained classification rates (we consider the averaged class specific rates) computed for increasing tree depths Z. The evaluation is performed for tree numbers of $T = 5$ (left) and $T = 10$ (right).

RGB	d_x,d_y	d_{xx},d_{yy}	Pixel Avg [%]	Class Avg [%]	Eval Time [s]
	X	X	73.82 ± 1.42	76.21 ± 0.67	1.89
X			77.88 ± 1.36	79.31 ± 0.82	1.58
X	X		86.48 ± 0.93	86.75 ± 1.04	1.72
X	X	X	87.54 ± 0.84	87.72 ± 1.28	2.20

Table 3.2: Pixel-wise interpretation results obtained for a set of 30 test images. Since there exists no suggested train/test data splitting, the rates are averaged over 20 runs with random sets of training and test images. We depict the accuracy for different combinations of integrated low-level feature cues.

tained quantitative results for different combinations of feature cues are summarized in Table 3.2. Figure 3.9 depicts some pixel-wise interpretation results using color, first and second order derivative filters. Note that even an initial scene geometry, *e.g.*, computed with SfM might help to further increase the classification accuracy.

Forest Size

The second experiment investigates the performance of using the Sigma Points representation together with RF classifiers by varying the forest sizes. The forest size is defined by the number of trees in the forest T and maximum depth Z of each tree. In this experiment we construct the Sigma Points by using low-level cues according to $F(x, y) = [R \;\; G \;\; B \;\; d_x \;\; d_y \;\;]^T$.

The classification accuracies are determined for forest sizes of $T = \{1 \ldots 20\}$ and

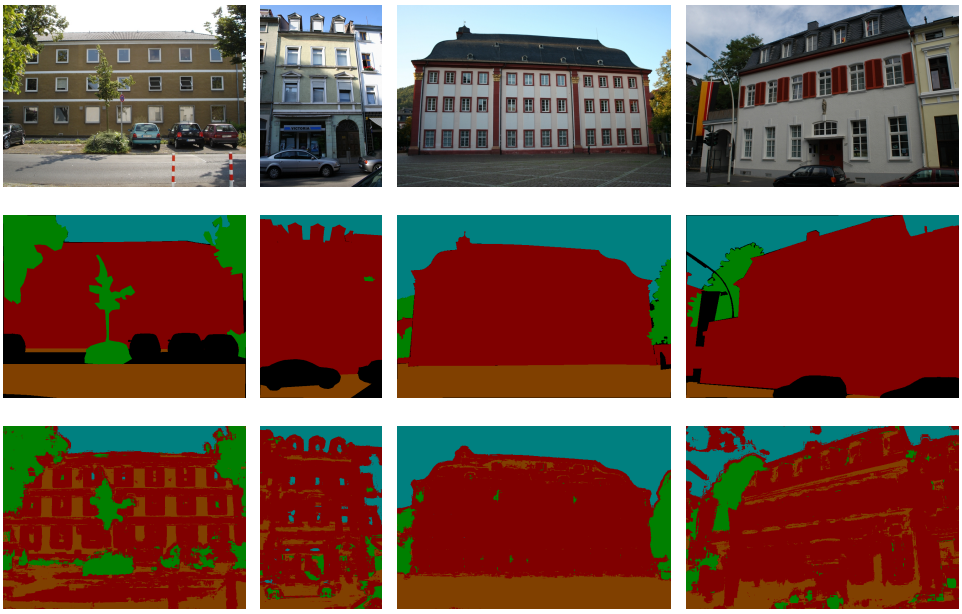


Figure 3.9: Pixel-wise semantic interpretation obtained for some eTRIMS facade images. The applied classification procedure segments the test images into four different object classes: *building* (red), *ground* (brown), *sky* (blue) and *vegetation* (green). The initial classification is obtained by integrating color, first and second order derivatives within the Sigma Points feature representation.

$Z = \{5 \dots 20\}$. We additionally record the required time for training and an averaged evaluation for a single test image. Due to randomness of the RF classifier we average the results over 5 runs. The obtained classification accuracies are given in Figure 3.10. The performance evaluation shows that a forest size of $T = 10$ and $Z = 15$ clearly reaches a saturation in terms of correctly classified pixel. Moreover, while the required training time increases exponentially, one can observe that the evaluation time is linear with the forest size since a feature vector evaluation is accomplished with $T \times Z$ binary tests.

3.8.3 Initial Interpretation of Benchmark Images

Considering the conduced experiments, a forest size of $T = 10$ and $Z = 15$ and an integration of RGB color/first order derivatives have proven to give a reliable initial semantic interpretation of the MSRCv2 test images. As a next step we use this setup to perform a semantic interpretation on all benchmark datasets. The obtained performance measurements are summarized in Table 3.3. One can notice that an initial interpretation of normally-sized images (circa 0.5 MPixels) takes in the range of 2 seconds. A detailed class-specific evaluation for the two MSRC image collections is given in Table 3.4,

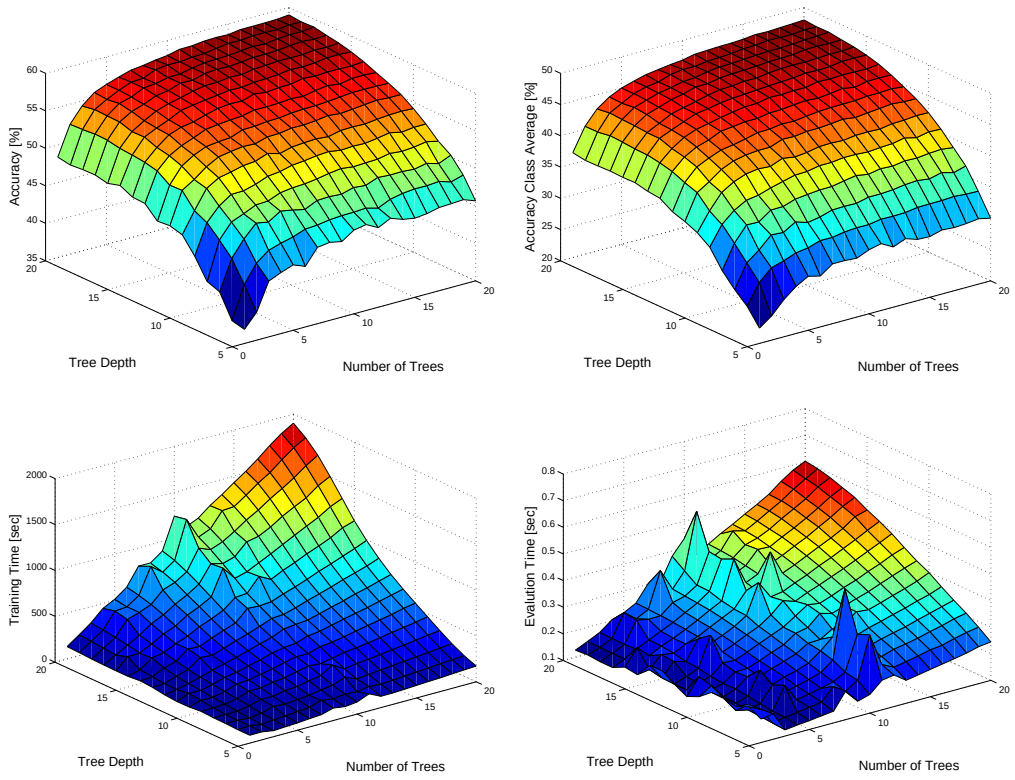


Figure 3.10: Computed classification accuracies depending on the forest size. The first row depicts the classification accuracy computed over all pixel (left) and averaged for each object class (right). A saturation of the classification accuracy is obtained for a forest size of $T = 10$ and $Z = 15$. The recording of the training (left) and evaluation time is given in the second row (right). The evaluation time increases linearly with the forest size and varies between 0.3 and 0.8 seconds for a reliable classification of a MSRC image. The training process takes about 20 minutes on a standard hardware.

where we can observe that our classification workflow successfully distinguish between *stuff* object classes, such as *grass*, *tree*, *water* and *road*, by solely using color and basic texture information. The interpretation of pixels, occupied by man-made objects, like buildings and cars, particularly fails due to ambiguous image regions (non-textured areas) and missing context information.

Figure 3.11 shows some interpretation results computed for MSRCv2 test images. From the visual results, it can be clearly seen that the proposed classification workflow does not preserve real object boundaries since the Sigma Points are compute within rectangular image patches. We therefore conduct additional experiments to overcome the problem of missing object delineation.

Dataset	# Classes	# Pixels [Pixel]	Pixel Avg [%]	Class Avg [%]	Eval Time/Image [s]
eTRIMS	4	396058	86.48	86.75	1.87
MSRCv1	9	68304	71.65	70.70	0.26
MSRCv2	21	69578	56.67	45.64	0.29
VOC2007	21	177625	17.89	16.80	0.70

Table 3.3: Performance evaluation of the proposed interpretation workflow. For each dataset separately, classifiers are trained and evaluated with extracted Sigma Points features. For a quantitative evaluation we use a forest size of $T = 10$ and $Z = 15$ and an integration of RGB color and first order derivatives. Although the rates for VOC2007 seem very low (16.8 %), the averaged accuracy is considerably higher than random chance (4.76 %).

3.8.4 Introducing Segmentation and Refined Labeling

In this section we investigate how unsupervised segmentation and the refined labeling influence the classification accuracy in terms of correctly classified pixels. In this thesis we solely concentrate on single image partitions, however we expect that multiple segmentation will further improve the accuracy with an increased computational complexity. In the following we mainly focus on the two different concepts to improve the initial image interpretation with respect to real object boundaries.

The first concept (we refer to it as *SuperPixel*) relies on a direct computation of statistical Sigma Points features for each super-pixel, provided by a Quickshift [Vedaldi and Soatto, 2008] image segmentation. It is obvious that the integral images cannot be used for this concept. Extracted super-pixel feature instances, together with their class labels (taken from the available ground truth maps), are learned and evaluated with the RF classifier. This concept thus involves a conservative super-pixel segmentation for the training process, but feature extraction and classification can be performed on reduced image data.

The second method (in the following we refer to it as *SpatialSupport*) is based on extracting features within rectangular patches located around a pixel. In this case integral structures permit an efficient construction of the feature representation. We train a classifier, where each obtained feature vector is assigned the most frequent class label within the rectangular patch analyzing the provided ground truth. Contrary to *SuperPixel*, where the classifier directly yields a class distribution for an entire segment, we treat a super-pixel more as a region providing important spatial support. At runtime, the classifier computes confidences for each pixel. An aggregation of confidences within a super-pixel then results in an averaged class distribution for each segment. Note that this method does not require any segmentation during training.

Object Class	MSRCv1 [%]	MSRCv2 [%]
building	**43.30**	**19.91**
grass	**85.66**	**88.78**
tree	**81.36**	**74.43**
cow	67.61	49.38
sky	86.71	67.78
aeroplane	54.26	50.86
face	68.61	58.29
car	64.98	32.66
bicycle	83.51	64.24
water	n.a	**53.93**
sheep	n.a	53.51
flower	n.a	60.61
sign	n.a	31.60
bird	n.a	13.74
book	n.a	45.69
chair	n.a	19.68
road	n.a	**53.39**
cat	n.a	35.49
dog	n.a	26.31
body	n.a	19.57
boat	n.a	20.17
chance	11.11	4.76
average	70.70	45.64

Table 3.4: Computed classification results individually obtained for the available object classes. The rates distinctly exceed the values of random chance. The obtained rates for relevant *stuff* object classes, that frequently occur in aerial images (*tree*, *grass*, *road*, *water*, *building*) are largely higher than the averaged classification rates and random chance, respectively.

We compute the Sigma Points, describing a super-pixel or an image patch (21×21 pixels), for RGB colors and the first order derivatives yielding a feature vector with a dimension of 55 for $D = 5$. As usual we train an RF with $T = 10$ trees, each with a maximum depth of $Z = 15$.

In order to generate the super-pixels, not limited to a size or number, we apply Quick-shift [Vedaldi and Soatto, 2008] to a five-dimensional vector consisting of image coor-

	MSRCv1		MSRCv2		VOC2007	
	Pixel Avg [%]	Class Avg [%]	Pixel Avg [%]	Class Avg [%]	Pixel Avg [%]	Class Avg [%]
Pixel level	71.65	70.70	55.86	44.76	17.06	16.08
SuperPixel	68.75	67.28	53.20	42.73	17.59	16.96
SpatialSupport	74.99	74.74	62.04	50.72	22.53	19.14
Refined	79.58	79.63	69.30	58.11	29.45	21.92
[Kluckner et al., 2009b]	86.80	81.8	73.7	61.8	29.1	21.45
[Schroff et al., 2008]	87.20	n.a.	71.7	n.a.	n.a.	n.a.
[Pantofaru et al., 2008]	n.a	n.a.	74.3	60.3	n.a.	n.a.
[Gould et al., 2008]	88.5	n.a.	76.5	n.a.	n.a.	n.a.

Table 3.5: A comparison of obtained classification rates for three benchmark datasets (MSRCv1,MSRCv2,VOC2007). The results are evaluated at the pixel level and illustrate the performance of the different stages of our approach: Raw pixel and super-pixel classification, introducing spatial support using the super-pixels, and the refinement step with the CRF stage ($\lambda = 2.0$), defined on the super-pixel adjacency graph.

dinates and CIELab color values. The parameters for Quickshift are set to $\sigma = 2.0$ and $\tau = 8.0$. The CRF-based refinement step is performed on the super-pixel graph by using $\lambda = 2.0$.

Note that the class distributions for the super-pixels are computed by using the methods *SpatialSupport*. Table 3.5 compares the initial results for the pixel-wise interpretation, both concepts *SuperPixel* and *SpatialSupport*, and the refined labeling to the rates reported for state-of-the-art methods. We again report both the overall per-pixel classification rate (*i.e.*, the accuracy of all pixels correctly classified) and the average of class specific per-pixel percentages, that gives a more significant measurement due to varying quantity of labeled pixels for each class.

Interestingly, the initial classifications, performed at the pixel level can be improved significantly by exploiting super-pixels as spatial support for all three datasets, while the direct classification of feature instances that are extracted within super-pixels, obtains a slightly lower accuracy. We assume that an improved integration of local context information due to exceeding the object boundaries and high redundancy within a super-pixel obtained by classification at the pixel level, cause the significant increase of the overall number of correctly classified pixels.

The CRF stage further improves the classification rates yielding a consistent final labeling of the super-pixels. Please note, the results presented in this work do not consider,

e.g., global context information [Pantofaru et al., 2008] or location priors [Gould et al., 2008]. This context data would avoid impossible object constellations within an image, like that a cow stands on a table. In [Kluckner et al., 2009b] we additionally used multiple image segmentation and context information learned from the trainings data. Some visual results are depicted in Figure 3.11. Figure 3.12 highlights massive failure cases obtained for the MSRCv2 dataset.

An evaluation of computational costs at runtime shows that the generation of the super-pixels for a MSRC image consumes most of the time (approximately 680 ms) in both concepts, *SuperPixel* and *SpatialSupport*. Feature extraction, classification and refined labeling on a reduced image grid (100 ms) runs slightly faster than evaluating all pixels in the test image by using efficient integral structures (160 ms). A comparison shows that our suggested concepts take about 1 second per image, which is 6 times faster than the timings reported in [Shotton et al., 2007]. Taking into account the computational costs and classification accuracy we conclude that the method *SpatialSupport* performs better than *SuperPixel*. It is evident that the CRF stage could also be optimized for a four- or eight-connected neighborhood.

3.9 Experiments on Aerial Images

So far, we have shown that our interpretation workflow, based on the Sigma Points feature representation and efficient RF classifiers, can be successfully applied to standard benchmark image collections. In a next step the proposed method is demonstrated on real-world aerial imagery, where we integrate appearance channels and 3D information to obtain a interpretation into multiple object classes. In this chapter we evaluate the proposed classification solely on single large-scale aerial images that are acquired from different viewpoints with high overlap.

In order to demonstrate the influence of the available feature cues (*i.e.*, color, edge responses and elevation measurements) we compute pixel-wise interpretation results for different combinations integrated within the Sigma Points representation.

In a first experiment we focus on the interpretation of single aerial images into two classes, where we discriminate building structures from the background. Note that our approach can be easily extended to handle other challenging binary classification tasks, *e.g.*, distinguishing between vegetation/non-vegetation or extracting water areas from the background.

A second experiment investigates the performance of the interpretation by using an extended set of objects classes (*building*, *tree*, *water*, *grass* and *street*), where we entirely segment each aerial image into those five object classes.

Due to spectral differences, varying pixel sizes and changing landscapes we indi-

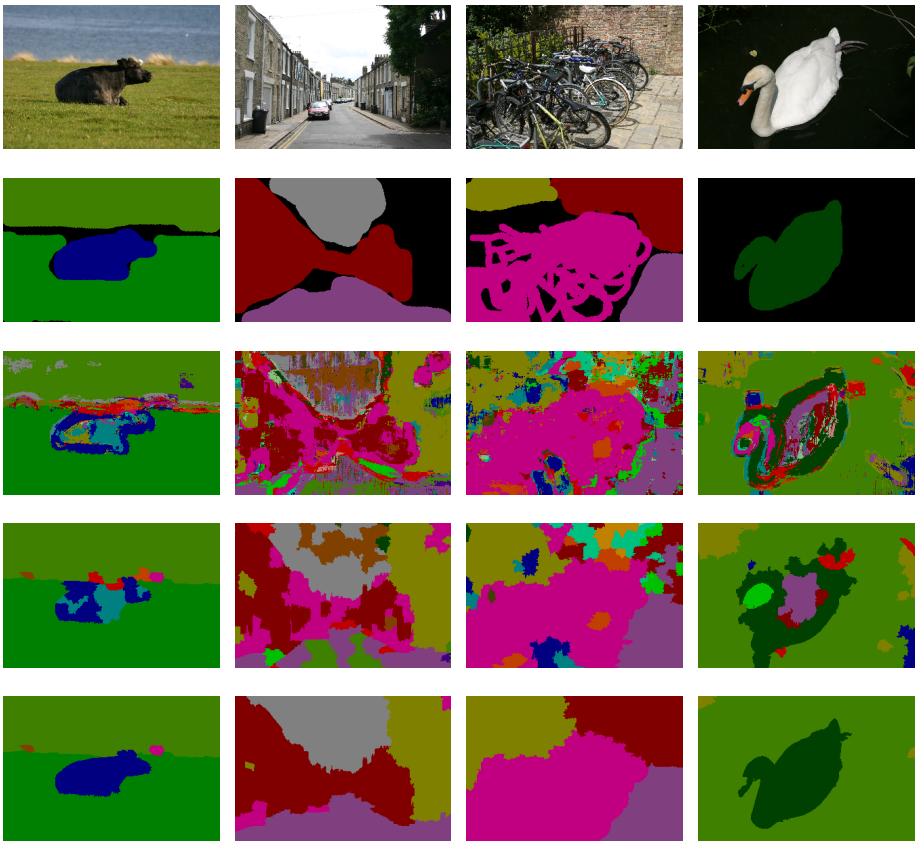


Figure 3.11: Some visual results selected from the MSRCv2 database including 21 object classes. The first and second row shows the original color images and the ground truth annotation, respectively. The initial results obtained by the RF classification are given in the third row. The fourth row shows the semantic interpretation results using a grouping of confidences within super-pixels (*SpatialSupport*). The CRF-refined final interpretations are depicted in the last row.

vidually construct the required classifiers for each aerial project (*i.e.*, *Dallas*, *Graz* and *San Francisco*) by using hand-labeled training data. For each dataset separately we thus have to annotate images providing the training labels at the pixel level. Moreover, non-overlapping ground truth maps are additionally generated for the evaluation procedure.

3.9.1 Manually Labeling

Since GIS information, in form of cadastral maps, and expert-labeled data are not available or out of date we have to generate ground truth maps in a rigorous process. While the large amount of overlapping aerial images is relatively easy to acquire, the annotations for

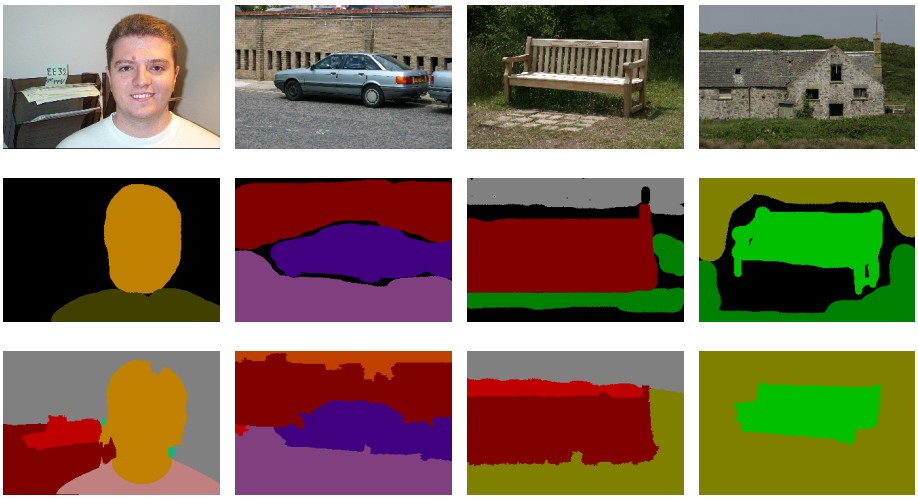


Figure 3.12: Some obvious failure cases selected from the MSRCv2 database including 21 object classes. Again, the first and second row show the original color images and the ground truth, respectively. The first three columns depict massively wrongly classified regions in the refined labeling (third row). In these cases (column 1-3), context information in form of probable class constellation within an image (a face appears above a body, boats and planes are generally not mounted on buildings), would significantly improve the final interpretation. In case of the last column, additional 3D information could help to enhance the final classification.

the buildings, water or grass areas, trees and street layer, are generally not available. Due to spectral differences and huge variations in the landscape and the appearance, an enormous amount of labeled training data is required for each aerial project. The coverage of extended sets of object classes particularly needs a tedious and sophisticated manual labeling effort.

Therefore, we have developed an approach to extract representative training labels. In a first step the corresponding camera centers get clustered spatially in order to select a representative collection of perspective images for the ground truth generation. The number of expected clusters is chosen according to the dimension of the described bounding box in the world coordinate system. The cluster centers then determine the nearest neighbors as probable candidates for the labeling procedure. Note that this concept only selects a basic set of potential images. A trivial manual interaction step involves an additional visual selection to gather special observations for the training process.
According to the quantity, the assembly and the shown landscape characteristics of the available images we then generate corresponding ground truth maps for each aerial project in a second step. It is obvious that the generation of training labels demands attention with

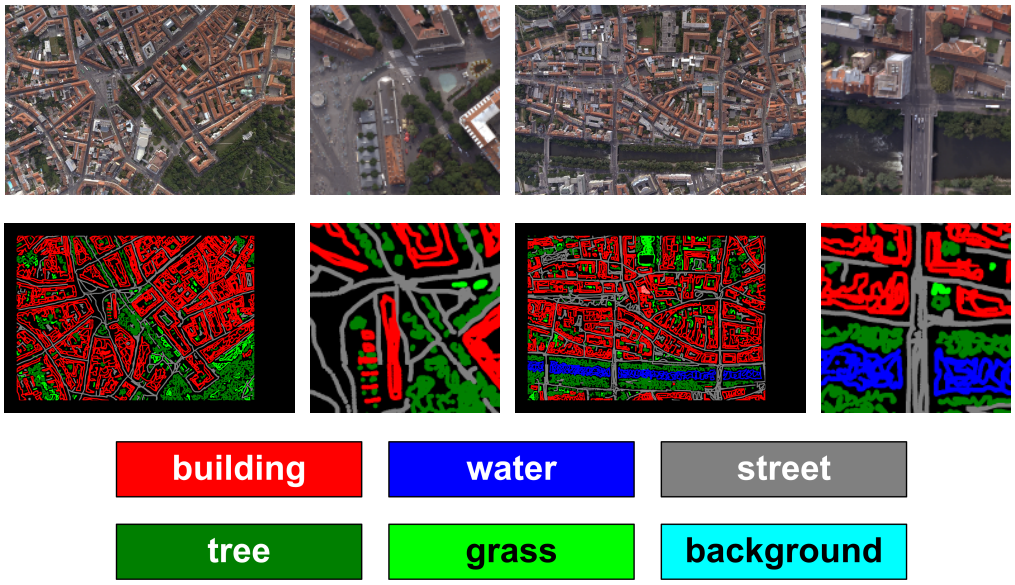


Figure 3.13: Manually labeled ground truth information for *Graz*. The objects in the images are annotated by using brush strokes. We commonly assign labels for the classes *building*, *water*, *grass*, *tree* and *street*. The color coding for the object classes is illustrated in the last row.

respect to the mapped objects. For instance, one has to distinguish between swimming pools and natural water areas. Thus, we rather assign pools a *building* class label than a *water* label. Moreover, it is not trivial to assign the training labels for elevated circulation spaces, like bridges or overpasses. In these ambiguous cases no labels are assigned to those image regions. In addition, street-related objects, such as cars, trucks and trams, are assigned with the *street* label. Please note that an interpretation at a country-scale requires additional object classes, *e.g.*, for the discrimination of different types of agriculturally used areas or mountains.

For the experimental evaluation quantities of 8 (*Dallas*), 15 (*Graz*) and 13 (*San Francisco*) images have been largely labeled to provide the ground truth information. The number of labeled pixels for each aerial project varies in the range of 300 MPixels to 350 MPixels. Figure 3.13 shows ground truth labels generated for images of *Graz*.

3.9.2 Binary Building Classification

To investigate the influence of the available low-level feature cues we collect Sigma Points vectors for different combinations of input sources. For the binary classification procedure we utilize the color images and elevations measurements, resulting from the DSM and the DTM. In addition, we again apply Sobel filter masks to gray-value converted color images

to compute elementary texture information. In this experiment we compute binary building classification results for combinations of elevation measurements, color, color/edge responses, and color/edge responses/elevation measurements. The feature vector then consists of 78 attributes, if RGB color, edge responses and elevation measurements are combined by using the Sigma Points representation. The dimension of the spatial neighborhood is set according to the provided datasets GSDs with 9×9 pixels (*Dallas*, *San Francisco*) and 13×13 pixels (*Graz*), respectively. During the training and the evaluation procedure we collect the feature representation at a full image resolution (11500×7500 pixels). For each dataset separately, RF classifiers with $T = 8$ trees and maximum depths of $Z = 15$ are trained in a first step. Note that the size of the forests has given a reliable trade-off between computation time and accuracy. Due to a significant reduction of expensive computation time, the trained RFs are evaluated at each pixel location using a third of the full image resolution. The final classification results, composed of the class-specific confidence maps, are then up-sampled to form a dimension-corresponding interpretation result.

The building classification rates obtained for the datasets *Dallas*, *Graz* and *San Francisco* are summarized in Table 3.6. A mix of color, edge responses and 3D information results in averaged rates of 92% (*Dallas*), 96% (*Graz*), and 90% *San Francisco*. For instance, using solely color yields low classification accuracies of 83% (*Dallas*), 83% (*Graz*), and 79% (*San Francisco*). In Figure 3.14 a building classification for *San Francisco* is shown by using different combinations of available input sources. It is evident that color only cannot successfully separate all building structures from the background since some image regions are massively covered by shadows. In these cases even for humans it is hard to distinguish between *building* and *background*. Figure 3.15 illustrates some building classifications obtained for *Graz* scenes. The tight combination of RGB color, first order derivatives and elevation measurements obviously yields the best interpretation results. Furthermore, one can notice that even the challenging bluish building structure (*Kunsthaus*) can be successfully separated from the background.

3.9.3 Semantic Interpretation into Five Object Classes

As a next step we treat the semantic interpretation as a five-class segmentation problem, where each pixel of the aerial images is assigned a label representing one of the classes for *building*, *tree*, *water*, *grass* or *street*. Similar to the building classification we individually train RF classifiers ($T = 8$ trees, depth $Z = 15$) for different combinations of feature cues.

The size of the considered neighborhood is again set according to 9×9 pixels for *Dallas* and *San Francisco*) and 13×13 pixels for *Graz*. Table 3.7 summarizes the ob-

Dallas					
RGB	Texture	Height	Pixel Avg [%]	*building* [%]	*background* [%]
		X	65.01	69.45	60.57
X			82.65	89.47	75.85
X	X		84.96	90.18	79.74
X	X	X	**91.46**	85.37	97.53

Graz					
RGB	Texture	Height	Pixel Avg [%]	*building* [%]	*background* [%]
		X	62.04	66.59	57.68
X			82.68	82.78	82.58
X	X		87.20	83.65	90.60
X	X	X	**96.05**	94.04	97.98

San Francisco					
RGB	Texture	Height	Pixel Avg [%]	*building* [%]	*background* [%]
		X	57.68	72.33	42.70
X			74.85	83.29	66.23
X	X		78.67	82.42	74.85
X	X	X	**90.44**	87.02	93.92

Table 3.6: Binary building classification results obtained for the three aerial datasets. The rates are given for different feature cues and combinations in terms of correctly assigned labels. It can be clearly seen that an integration of appearance and 3D information significantly improves the interpretation at the pixel level.

tained classification results in terms of correctly classified pixels. We can observed that a combination of color, edge responses and 3D information significantly improves the classification accuracy. Moreover, it is interesting to notice that especially height only successfully separates the *water* object class from the remaining classes. We assume that the high reconstruction errors, caused by motion, yield high variances in the statistics estimation. For the RF classifier it is thus rather easy to select discriminative attributes in the provided Sigma Points feature vector.

In Figure 3.16 computed confusion matrices are shown for color as feature cue, and the combination of color, basic texture and 3D height information. Figure 3.17 shows

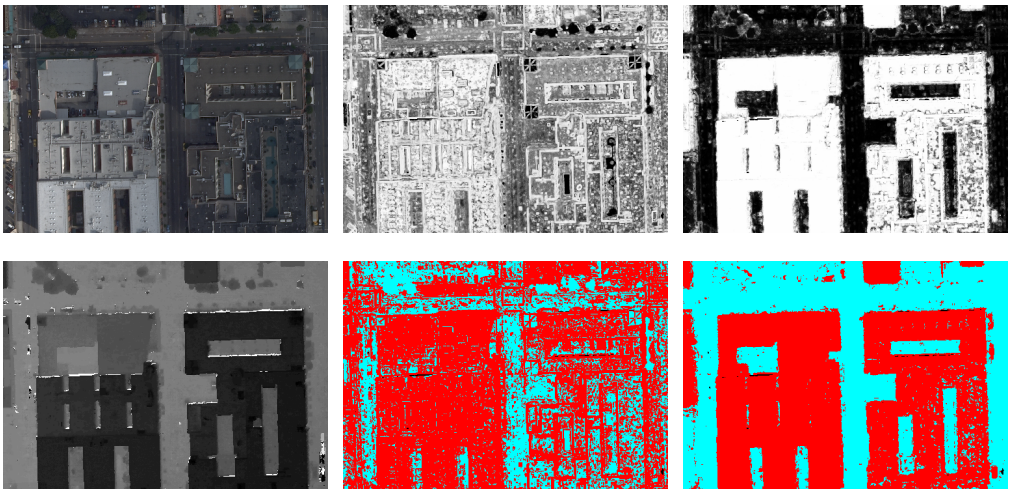


Figure 3.14: A building interpretation result for a scene of *San Francisco* using only color and the entire set of available feature cues (color, edge responses, elevation measurements). Color and available 3D information are shown in the first column. In the middle, the obtained building classification is given for the use of RGB color information (we depict the raw building confidences and the dominant object class. The last column illustrates the corresponding results obtained for color, basic texture and elevation measurements. Red colors denote the *building* class, while cyan describes *background* areas.

a semantic interpretation into the five object classes by using different combinations of appearance and elevation measurements. One can observe that the use of height only produce many water areas. Similar to moving water areas, regions occupied by specular or occluded facades result in uncertain range information. A tight integration of appearance and color information within the proposed interpretation workflow significantly overcomes these problems. Moreover, it is obvious that a regularization scheme would further improve the final assignment of the object classes. Figure 3.18 shows an interpretation results for overlapping images of *Dallas*. These images clearly show that frequently occurring crossovers are largely assigned with the *building* class. Note that a labeling of these structures as street would considerably worsen the final classification accuracy. Thus, context information in form of web-based GIS data may be exploited to adapt these problems. In Figure 3.19 the raw confidences for each object class, directly evaluated with RF classifiers, are shown for some images of *Graz*.

3.9.4 Improving the Initial Classification Result

Until now the semantic explanation of the aerial scenes has been performed at the pixel level without consideration of the spatial neighborhood information. As shown in Fig-

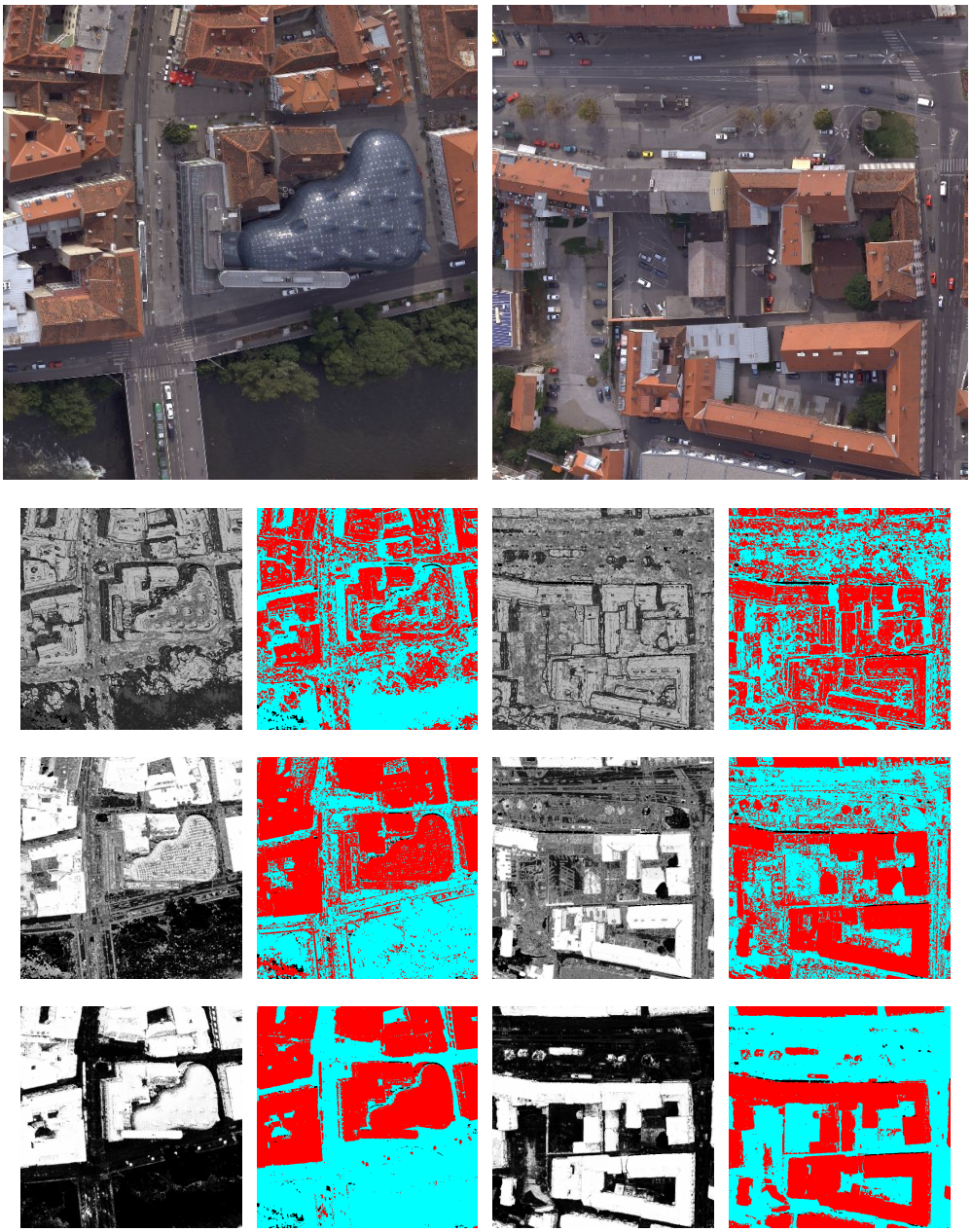


Figure 3.15: A building classification computed for two *Graz* scenes by using different combinations of low-level feature cues. The color information, represented in a perspective view is given in the first row. The second row depicts the raw confidences and the dominant object class for the use of color. A result by utilizing color and basic texture is shown in the third row. From the last row, it can be clearly seen that an integration of color, texture and elevation measurements yields the best interpretation result.

Dallas								
RGB	Texture	Height	Pixel Avg [%]	*building* [%]	*water* [%]	*grass* [%]	*tree* [%]	*street* [%]
		X	42.75	35.15	67.52	58.71	71.73	28.87
X			77.65	72.10	94.67	88.73	91.97	73.42
X	X		81.49	78.15	92.10	91.00	92.92	75.57
X	X	X	**89.36**	84.02	94.16	94.59	96.08	93.57

Graz								
RGB	Texture	Height	Pixel Avg [%]	*building* [%]	*water* [%]	*grass* [%]	*tree* [%]	*street* [%]
		X	40.05	38.36	80.68	38.97	31.76	44.03
X			73.68	64.25	77.99	61.42	86.56	85.49
X	X		79.53	71.73	88.40	66.97	93.14	86.36
X	X	X	**91.55**	93.23	96.11	75.56	92.91	89.81

San Francisco								
RGB	Texture	Height	Pixel Avg [%]	*building* [%]	*water* [%]	*grass* [%]	*tree* [%]	*street* [%]
		X	28.73	29.89	84.26	65.43	51.50	0.00
X			65.04	62.37	91.87	69.44	74.38	59.32
X	X		70.69	71.87	95.74	69.66	82.14	59.38
X	X	X	**82.98**	81.20	96.32	71.71	81.52	84.52

Table 3.7: The results obtained for a semantic interpretation into five object classes for the three aerial datasets. The rates are given for different feature cues and combinations in terms of correctly assigned labels.

ure 3.20, the interpretation shows a high granularity with respect to the dominant class label. Similar to the refinement step, that is applied to the benchmark images, super-pixel segmentation and the CRF stage are subsequently used to obtain a smooth and consistent class assignment. In this experiment we utilize a CRF stage defined over a four-connected neighborhood, and the one computed for a super-pixel generated adjacency graph. The optimized class assignments are summarized in Figure 3.20. For both optimization strategies the smoothing parameter is set to $\lambda = 10.0$. In addition, the available color image provides the weights for the graph edges, where we either calculate simple color distances between neighboring pixels or the deviation between the mean colors of computed super-pixels. We estimate the weighted color distance by using a exponential normalization

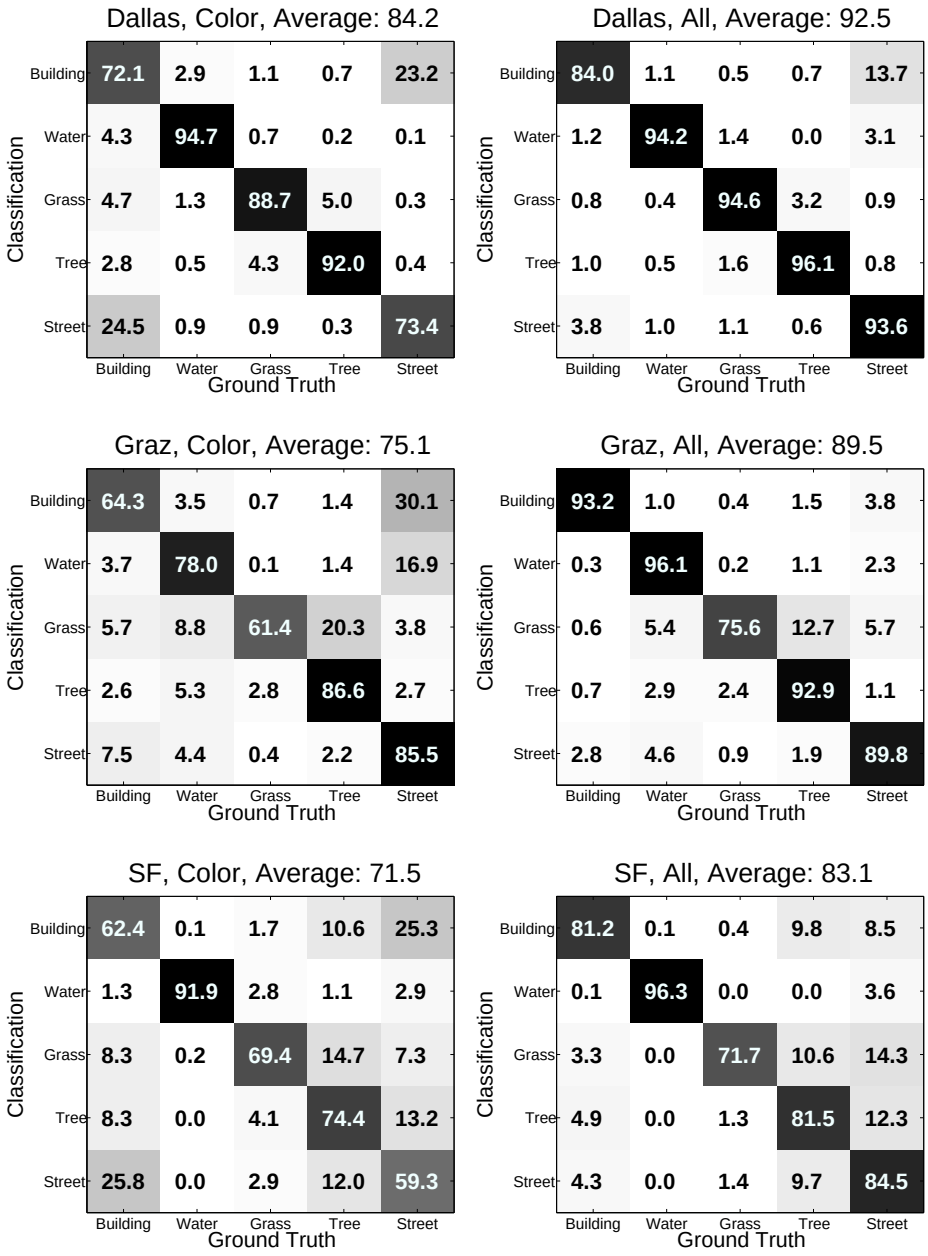


Figure 3.16: Computed confusion matrices for the three datasets. It is obvious that an integration of color, texture and 3D information (right column) drastically improves the classification accuracy rates compared to use color only (left colomn). This combination particularly benefits the discrimination between the *building* and the *street* class.

function with $\exp(-\kappa|\cdot|)$. The weights $\kappa = \{0.1 \ldots 10.0\}$ are adjusted according to a visual inspection of the results.

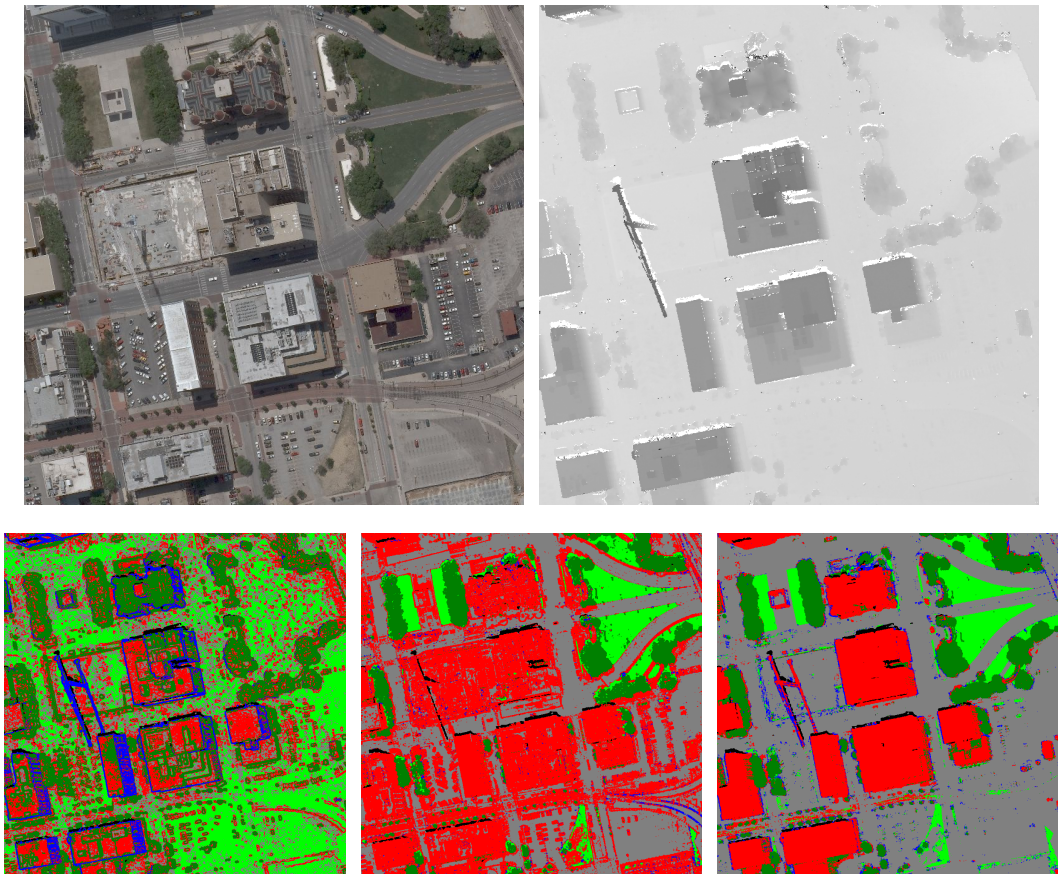


Figure 3.17: A five-class semantic interpretation result computed for a *Dallas* scene by utilizing appearance and elevation measurements. The first row presents both the original color image and the derived DSM. The second row shows the pixel-wise interpretation result obtained by using height only, color/texture, and color/texture/3D information. From a visual inspection one can notice that an integration of appearance and elevation yields the best results.

An optimized labeling is given in Figure 3.21 for a small *San Francisco* scene. In Figure 3.22 the original color image is overlaid with both a raw building classification and the optimized binary labeling. One can notice that the explanation of building structures, in particular the facade elements, can be improved by applying the CRF stage. However, although the refinement would partly improve the intermediate interpretation, we skip the optimization procedure within our workflow due to enormous computational effort. Refining an image with a dimension of 800×700 pixels takes approximately 5 seconds. Thus, an optimization of an aerial image at full resolution requires in the range of 15 minutes, which would scale to hours for an entire aerial project. We rather optimize the

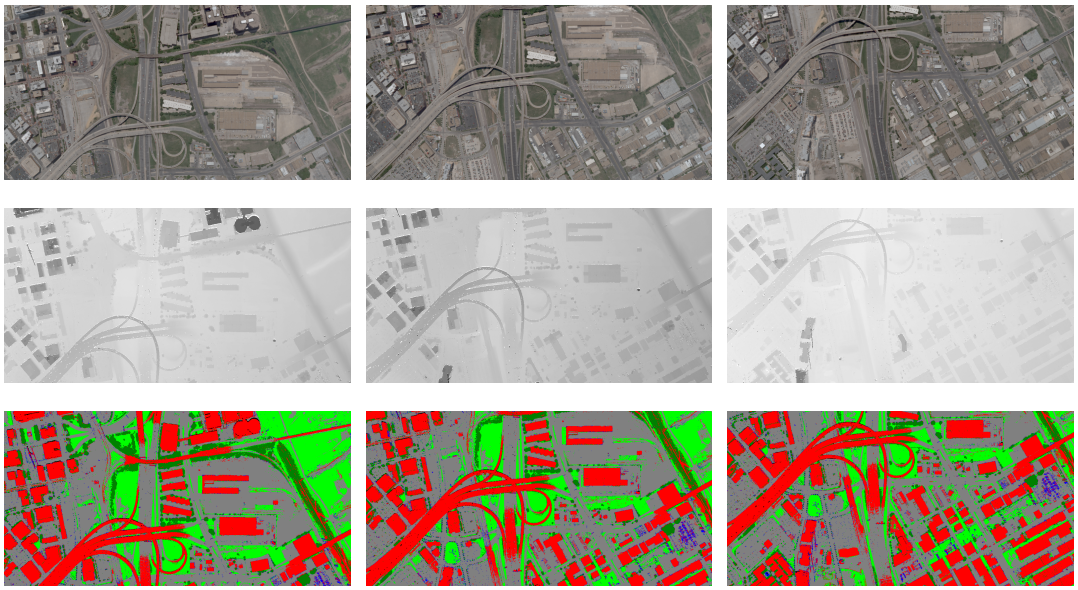


Figure 3.18: Initial semantic interpretation obtained for huge aerial images (11500×7500) showing the environment of *Dallas*. The classification procedure segments every test images into five classes by integrating color, first and second order derivatives within the Sigma Points feature representation. Note that a single image can be processed within few minutes.

labeling for the fused classification results within an orthographic view as shown in the next chapter.

3.10 Discussion and Summary

This chapter has extensively focused on initial semantic interpretation of individual (aerial) images by integrating multiple types of modalities, such as appearance, edge responses, and 3D height information. In this work the semantic classification, together with the available color and geometry information, is an indispensable part of a holistic description of the image content. We thus have presented the Sigma Points features, a novel representation in the context of semantic image classification. They are efficiently derived from statistical region descriptors and enable a low-dimensional, straightforward and tight integration of multiple low-level cues and also exploits the correlation between them. Due to the representation on a Euclidean vector space, the Sigma Points can be easily applied to fast multi-class randomized forest classifiers.

In the experimental evaluation, we have shown state-of-the-art performance on standard benchmark datasets, like the MSRC or the VOC2007 image collections, without uti-

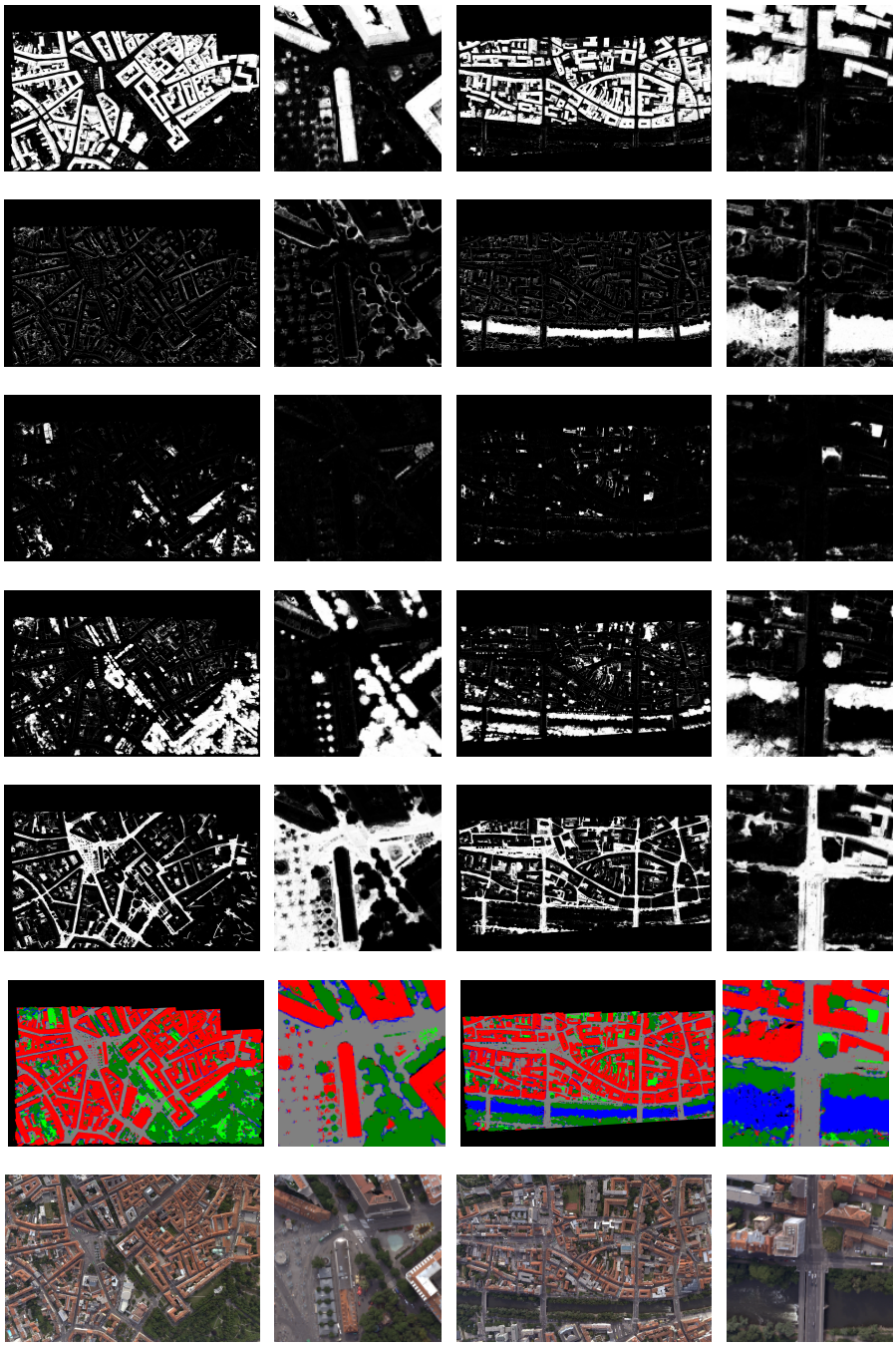


Figure 3.19: Initial semantic interpretation obtained for two images extracted from the aerial project of *Graz*. From the top to the bottom we depict the raw confidences for *building*, *water*, *grass*, *tree* and *street*, the dominant object class and the corresponding color image. We use appearance and elevation to compute the semantic interpretation.

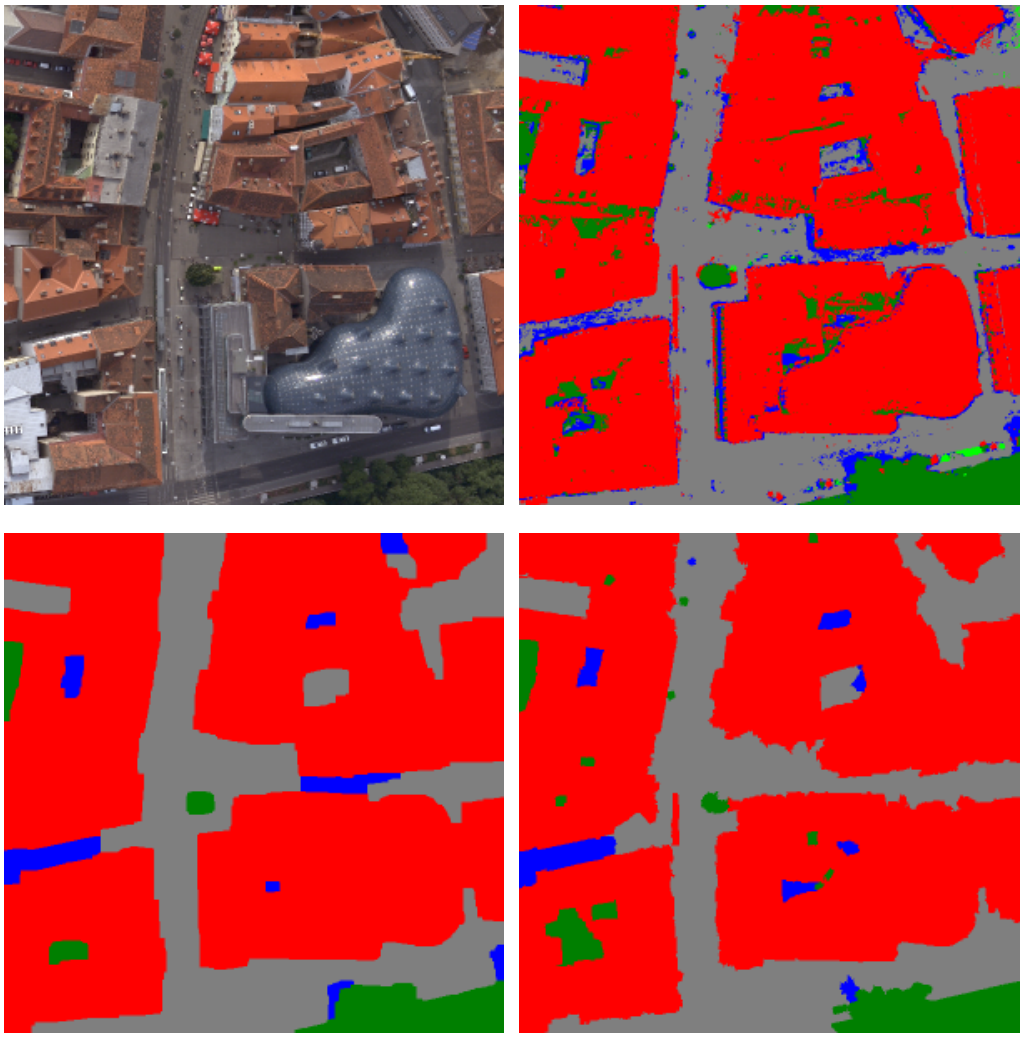


Figure 3.20: A refined labeling for a *Graz* scene (800×700 pixels). The first row shows the color image and the raw interpretation result, where the dominant object classes are evaluated at the pixel level. The second row depicts the optimized labeling obtained for the refinement on a four-connected and a super-pixel graph, respectively. While the super-pixel based optimization appears fringed, especially in facade regions, the image-graph refinement considerably discards small objects, like isolated trees and small courtyards. In addition, shadows cause large areas of wrongly assigned *water* labels. To reduce these problems, we propose to use a multi-view classification as a next step (see Chapter 4).

lizing derived location priors or global context cues, that handle valid class constellations within a test image. Moreover, we have demonstrated that an integration of unsupervised segmentation and energy minimization techniques improves the accuracy significantly by taking into account the initial pixel-wise classifications.

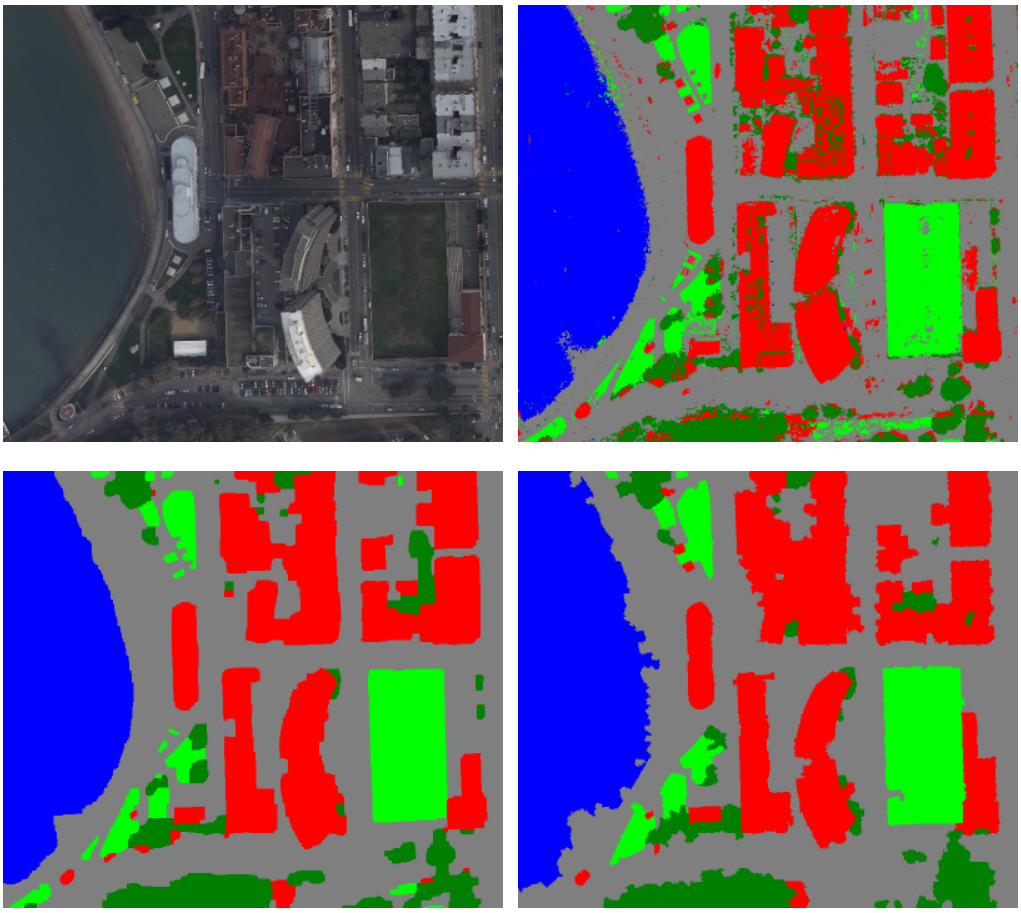


Figure 3.21: A refined labeling for *San Francisco* (800×700 pixels). The color image and the corresponding raw interpretation are depicted in the first row, while the second row shows an optimized class assignment using four-connected and a super-pixel graph, respectively.

In case of the aerial image interpretation we have shown that a combination of appearance and 3D height information is essential for an accurate explanation of each pixel. We have thus applied our method to the prominent task of building classification and to the multi-class problem, where we distinguish between objects representing the classes *building*, *water*, *grass*, *tree* and *street*. Our approach can be easily extended to handle additional object classes, such as different types of agriculturally used areas or mountains. The experiments have shown that the compact integration of appearance and 3D information yields an initial interpretation accuracy of more than 90% for the building classification and 81% to 93% for the pixel-wise multi-class explanation.

Until now we have considered the initial interpretation of single aerial images. Applying the interpretation to every image in the dataset provides us with a highly redundant

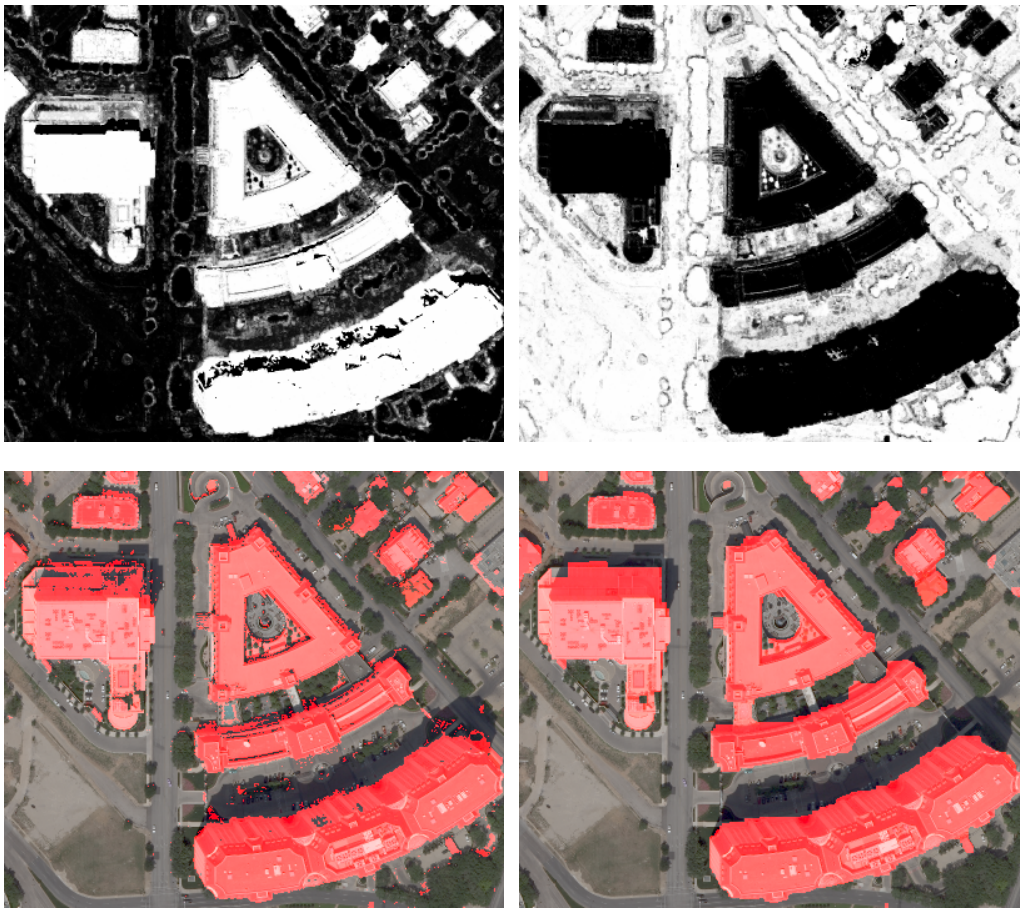


Figure 3.22: A building classification result computed for a *Dallas* scene. From left to right: the building confidences, the background probabilities, the color image overlaid with the raw classification result, and the color image masked with the refined interpretation using a four-connected image grid.

holistic collection of appearance, surface models and semantic interpretation. In fact, we are rather interested in generating a high-quality land-cover map within a meaningful description that makes extensively use of redundant image information. In aerial imagery, this representation is the orthographic view. Due to the availability of the scene geometry, the highly redundant results can be combined within this orthographic view. In the next chapter we therefore discuss the mapping and the multi-view fusion of color, height and semantic interpretation.

Chapter 4

From 3D to the Fusion of Redundant Pixel Observations

This chapter addresses the problem of aerial data fusion with many redundant observations for color, 3D information and the derived semantic interpretation. In particular, we present approaches for image fusion within an orthographic view, since this is a commonly used representation an aerial imagery workflow to integrate overlapping yet perspective images. Hence, the obtained fusion results for the available modalities, given in corresponding perspective view, can then be seen as a compact holistic description of the observed scene within the common view.

As a first step, the available range images and the camera data are extensively used to align multiple source modalities, such as appearance, height and classification results. For each modality individually, we present efficient methods, defined over redundant input images, to compute an improved fusion result. Next, we propose a variational formulation for the integration of redundant 3D height data, which then forms a consistent surface model. Furthermore, this formulation is extended with a wavelet regularization in order to enable a natural-appearing recovery of fine details in the images by performing joint inpainting and denoising from a given set of redundant color observations. Besides, we discuss an exemplar-based inpainting technique for an integrated removal of non-stationary objects, like moving cars. Finally, we additionally introduce a variational refinement step to make the semantic interpretation, aggregated from initial multi-view classification consistent and smooth with respect to real object boundaries. In the experimental section we show that the redundancy significantly helps to increase the image quality and the classification accuracy.

4.1 Introduction

The image fusion from digital aerial images poses a challenging problem for many reasons. In case of high-resolution color images, the fusion step must maintain fine details and complex textures. In addition, the fused images should provide natural appearance without outliers and erroneously assigned or missing regions. In contrast to a color image fusion, where object edges should appear naturally, the integration of redundant height information must preserve sharp edges and capture the 3D shape of objects in order to accurately describe and model, *e.g.*, the mapped rooftops. Moreover, the fusion has to handle large areas with missing or noisy data caused by moving cars, flowing water bodies or large areas of occlusion.

In general, image fusion integrates information of several images, taken from the same scene, in order to generate an improved image with respect to noise, outliers, illumination changes etc. Nevertheless, the efficient fusion of multiple observations for various modalities, ranging from scalar values [Zach et al., 2007, Carlavan et al., 2009], over vector-valued data, like color [Fitzgibbon et al., 2003, Agarwala et al., 2006, Strecha et al., 2008], to intermediate recognition results provided by various types of classifiers [Xiao and Quan, 2009, Leibe et al., 2007] is a hot topic in current research since scene information can be taken from different view points without additional costs.

In particular, modern aerial imaging technologies provide multi-spectral images, that map every visible spot of urban environments from many overlapping camera viewpoints. Typically, a point on ground is at least visible in 10 cameras. Figure 4.1 shows the resulting camera positions of the aerial project of *Graz*. Each camera location provides multi-spectral information about the observed scene.

The provided, highly redundant data (several observations for a considered point) enables efficient techniques for height field generation [Hirschmüller, 2006], but also methods for resolution and quality enhancement [Fitzgibbon et al., 2003, Agarwala et al., 2006, Strecha et al., 2008, Goldluecke and Cremers, 2009].

One the hand, initially derived range images, mainly computed in 2.5D from a stereo setup [Hirschmüller, 2006] or even from many input images [Irschara, 2011] can also be exploited to geometrically align data, such as the corresponding color information and semantic interpretation, within a common coordinate system. On the other hand, taking into account redundant observations of corresponding points in a common 3D world, the localization accuracy can be significantly improved using an integration of redundant range data for a full 3D reconstruction [Zach et al., 2007] or for a generation of improved 2.5D surface models [Zebedin, 2010].

In our approach we extensively utilize the alignment of many overlapping images within a common view. The available range images (see Chapter 2) are used to com-

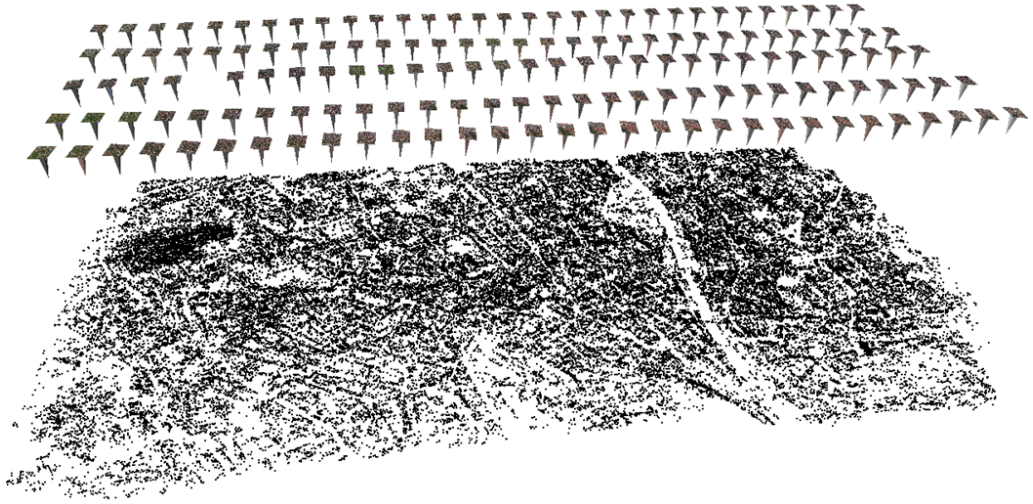


Figure 4.1: The aerial project of *Graz*. Each of the 155 camera locations provides multi-spectral information about the observed scene on ground from varying viewpoints. We fuse the perspective information of color, height and semantic interpretation into a common (orthographic) view.

pute pixel-wise geometric transformations between the original perspective images and an orthographic view, which is also related to novel view synthesis [Fitzgibbon et al., 2003, Woodford et al., 2007]. Since the range information is available for each image in the dataset, a transformation to a common view produce many redundant scene observations. Figure 4.2 depicts an aerial scene taken from different camera positions and a corresponding set of range images, that enable a geometric projection to a common view. Due to occlusions, specular and moving objects some of the range image tiles show large areas of missing information and erroneously assigned pixel information.

Thus, our task can also be interpreted as an information fusion from multiple input observations of the same scene by joint inpainting and denoising. While inpainting fills undefined areas, the denoising removes strong outliers and noise by exploiting the high redundancy in the input data. In our case, these outliers can be strong reconstruction errors included in the initial range maps, but also inconsistently transformed color values or wrongly assigned semantic class labels mainly caused by classification uncertainty or inaccurate object delineation. In this chapter we therefore concentrate on methods which account for these issues.

In literature, there exists a variety of methods to obtain improved fusion image results from multiple measurements. The methods differ in the used modalities, concepts and aims. Intuitively, considering scalar-valued modalities, like height and infrared data, a pixel-wise mean or even a median computation would produce a meaningful final result

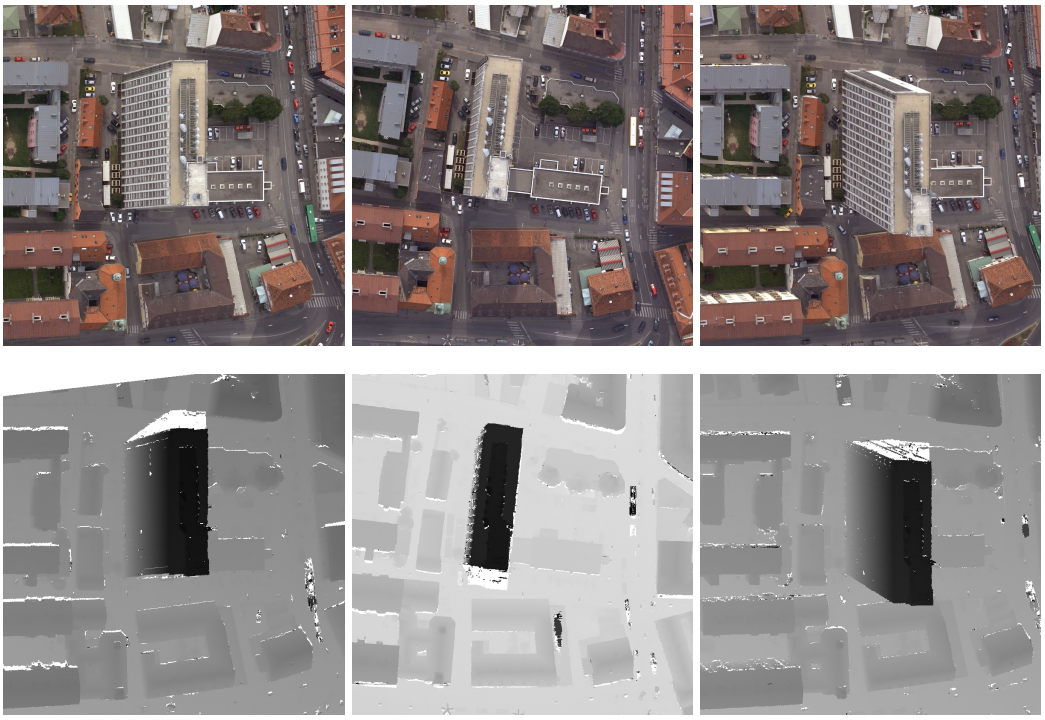


Figure 4.2: An aerial scene taken from different camera viewpoints and a corresponding set of computed range images, that enable a pixel-wise geometric transformations, together with the camera data, of image data to a common view. Occlusions, specular and moving objects cause large areas of missing information (white areas) and erroneously assigned pixel values.

with low computational complexity. In case of robustly fusing vector-valued data, such as the color information, a set of random projections of the entire vector onto 1D lines could be used to detect the median of high-dimensional information [Tukey, 1974]. Related, one could also use a clustering technique to find reliable color values for an improved texture generation [Zebedin et al., 2006]. In contrast to estimating a representative model of the given data, the classification result fusion is frequently performed by simply accumulating the provided confidences since the class confidences are in general normalized and produced from a single classifier.

Nevertheless, these fusion techniques provide low computational complexity, however, every pixel in the image domain is processed separately and, thus, the results are prone to be noisy and may contain outliers. To overcome these issues, energy minimization techniques with different types of regularizations are widely adopted to inverse problems [Tikhonov, 1943]. Those methods are based on the minimization of a given energy functional by considering both the deviation between the solution and an observed

input data and the smoothness of the solution, also referred to as regularization. In this chapter we investigate optimization schemes, formulated in the continuous domain to robustly integrate the available modalities. In particular, different types of regularizations are investigated.

In order to obtain fused information from multiple observations we first propose to use an extended formulation of the popular denoising model proposed by [Nikolova, 2004] to perform a tight integration of multiple intensity images, like color images or height fields. For clarity, we derive our model for scalar-valued images, however the formulation can be easily extended to vector-valued data. In addition, we show that this model is also useful to generate improved surface models and that a novel extension of the regularization term (we replace total-variation (TV) norm by an efficient wavelet based regularization) will lead to a natural-appearing fusion of multiple color images. Further, we again exploit the pixel-wise transformation to aggregate the computed class distributions, resulting from the initial semantic classification, into the common view. In order to regularize the aggregated interpretation result, we use a continuous formulation of the Potts model [Pock et al., 2009] to obtain a spatially coherent semantic labeling.

The experimental evaluation investigates the influence of redundancy on modalities, such as color and height, and presents qualitative and quantitative results. In addition, we show that the classification accuracy can be improved by collecting redundant probabilities for each object class from multiple view points. Moreover, we compare the refined labeling results obtained for different types of optimization strategies.

4.2 Introducing a Common View

As a first step, the required source images for modalities representing color, height and semantic interpretation need to be aligned within a common view. To obtain the required range data for each input image we use a dense matching similar to the method proposed in [Hirschmüller, 2006]. By taking into account the corresponding ranges images (see Figure 4.2) and the available camera data, each pixel in the images can be transformed individually to 3D world coordinates (the system can be either a local or a global one) forming a large cloud of points, where each point is then defined by a 3D coordinate, the color information and a class distribution, that encodes a probable object class assignment.

As commonly used in photogrammetry, we transform the data to orthographic view representation. Since we are interested in a large-scale holistic description of the scene within the image space, we introduce virtual orthographic cameras with a specified pixel resolution (we use a similar sampling distance as provided by the original images) in order to sample the information in a common 3D coordinate system. This enables a collecting of corresponding images tiles, out of the generated point cloud that is projected

to the ground plane (we simply set the height coordinate to a fixed value). Hence, a collection of multiple sampled points results in a highly redundant pool of image candidates for the different modalities. In Figure 4.3, the corresponding image patches, transformed to a common (orthographic) view, are given for color, height and semantic classification, where we only display the most dominant object class. Due to missing data in the individual height fields (*e.g.*caused by non-stationary objects or occlusions) the initial alignment causes undefined areas, artifacts or outliers in the novel view.

In the following various regularized approaches, individually adapted to the considered modality, are introduced to improve the final result with respect to these problems.

4.3 Fusion of Redundant Intensity Information

This section highlights our fusion model, that allows us to efficiently fuse redundant color and height field observations into a single, high-quality result. Note that an integration of multiple images at a city-scale demands fast and sophisticated methods, we thus make use of energy minimization techniques defined in the continuous domain. These methods provide a high parallelization capability and obtain in general a globally optimal solution. The basic concept of energy minimization methods is to formulate the solution of a given problem as an optimum of an functional, by taking into account of both the distance of the solution to the observed data and the smoothness itself. First, we briefly discuss related work. Then, we derive our fusion model, which is based on TV-L^1 denoising model [Nikolova, 2004], capable to handle multiple input observations. In order to exploit the particular characteristics of the involved modalities we individually adapt the fusion model for color and height information. In case of color we replace the TV-norm by a wavelet-based regularization, capable to preserve fine image structures and produces a natural fusion result.

4.3.1 Background

The challenging task of reconstructing an original image from given (noisy) observations is known to be an inverse ill-posed problem. In order to solve such optimization problems, additional assumptions, such as the smoothness of the solution, have to be considered. The optimization problem can be generally formulated as the minimization of the functional

$$\min_u \{ \mathcal{R}(u) + \lambda \mathcal{D}(u, f) \}, \tag{4.1}$$

where u is the sought solution, the function $\mathcal{R}(u)$ denotes the regularization, which forces a smooth solution. Depending on the problem $\mathcal{R}(u)$ can be defined for different types

Figure 4.3: Perspective scene observations transformed to a common orthographic view by using the corresponding range image information (see Figure 4.2). Our approaches take redundant observations for color (first row), height (second row) and semantic classification (last row) as input data and generate an improved fused image without undefined areas and outliers. Note that the individual observations include many undefined areas (white pixels) caused by occlusions, sampling effects and non-stationary objects.

of potential functions. The term $\mathcal{D}(u, f)$ expresses the data term, which measures the deviation between solution and one or even more observations f of the data. The scalar value λ controls the trade-off between data fidelity and the smoothness of the solution.

Although fast mean or median computation over multiple pixel observations will sup-

press noisy or undefined areas, each pixel in the image is treated independently (the image reconstruction is thus performed without spatial regularization). A variety of algorithms for fusion of redundant information is based on image priors [Fitzgibbon et al., 2003, Woodford et al., 2007], image transforms [Pajares and de la Cruz, 2004], *Markov* random field optimization procedures [Agarwala et al., 2006] and generative models [Strecha et al., 2008]. Variational formulations are well-suited for finding smooth and consistent solutions of the inverse problem by exploiting different types of regularizations [Tikhonov, 1943, Rudin et al., 1992, Nikolova, 2004, Carlavan et al., 2009, Zach et al., 2007].

The quadratic model [Tikhonov, 1943] uses the L^2 norm for regularization, however, causes smoothed edges. Introducing a TV-norm instead leads to the edge preserving denoising model proposed by Rudin, Osher and Fatemi (ROF) [Rudin et al., 1992]. The authors in [Nikolova, 2004] proposed to also use a L^1 norm in the data term in order to estimate the deviation between sought solution and input observation. Thus, compared to the ROF model, the resulting TV-L^1 model is more effective in the removal of speckle noise [Nikolova, 2004] and for shape denoising due to the contrast invariance provided by the L^1 norm [Nikolova et al., 2006]. Zach *et al.* [Zach et al., 2007] applied the TV-L^1 to robust range image integration from multiple views. Although TV-based methods are well-suited for tasks, like range data integration, in texture inpainting the regularization produces results that look unnatural near recovered edges, since regions are largely recovered by piece-wise constant solutions. In the last years a lot of effort has been put into the development of more suitable image priors.

To overcome the problem of the so called cartoon-like appearance caused by the TV-regularization, natural image priors based on multi-level transforms, like wavelets [Selesnick et al., 2005, Portilla and Simoncelli, 2000, Fadili et al., 2009] or curvelets [Candés et al., 2006], can be used within the optimization model [Starck et al., 2004, Carlavan et al., 2009]. These transforms provide a compact yet sparse image representation obtained with low computational costs. Inspired by [Carlavan et al., 2009], we utilize a wavelet transform for natural regularization within our proposed variational fusion framework capable to handle multiple input observations for color.

In the following the fusion model, defined over multiple input observations, is derived from well-established TV-L^1 denoising model. In a first step we use the model to integrate several observations for height information. Secondly, for the color image fusion the regularization is extended with a linear wavelet transform, that provide improved image priors yielding natural appearing fusion results.

4.3.2 The Proposed Model

We consider a discrete image domain Ω as a regular grid of size $W \times H$ pixels with $\Omega = \{(i,j) : 1 \leq i \leq W, 1 \leq j \leq H\}$, where the tupel (i,j) denotes a pixel position in the domain Ω.

Our fusion model, taking into account multiple input observations can be seen as an extension of the TV-L^1 denoising model proposed by Nikolova [Nikolova, 2004]. Especially the TV-L^1 denoising model will enable a reliable removal of outliers, that are caused by the geometry-driven alignment (we expect that the range images contain some errors). In the discrete setting the minimization problem of the common TV-L^1 model for an image domain Ω is then formulated as

$$\min_{u \in X} \left\{ \|\nabla u\|_1 + \lambda \sum_{i,j \in \Omega} |u_{i,j} - f_{i,j}| \right\}, \tag{4.2}$$

where $X = \mathbb{R}^{WH}$ is a finite-dimensional vector space provided with a scalar product $\langle u, v \rangle_X = \sum_{i,j} u_{i,j} v_{i,j}$, $u, v \in X$. The first term $\|\nabla u\|_1$ denotes the TV of the sought solution u and reflects the regularization in terms of a smooth solution. The gradient ∇u normally results from finite difference calculations with Neumann boundary conditions and is defined in the vector space $\mathbb{R}^{WH} \times \mathbb{R}^{WH}$. Moreover we also need the divergence operator $div : \mathbb{R}^{WH} \times \mathbb{R}^{WH} \rightarrow \mathbb{R}^{WH}$. This operator defines the adjoint of the gradient operator with $div = -\nabla^*$. The second term accounts for the summed errors between u and the (noisy) input data f. The scalar λ controls the fidelity between data fitting and regularization, whereas a low value of λ leads to smoother solutions. In following we derive our model for the task of image fusion from multiple observations.

As a first modification of the TV-L^1 model defined in (4.2), we extend the convex minimization problem to handle a set of K input observations $(f_1, \ldots, f_K)$. Introducing multiple input images can be accomplished by summing the deviations between the sought solution u and available observations $f_k,\ k = \{1 \ldots K\}$ according to

$$\min_{u \in X} \left\{ \|\nabla u\|_1 + \lambda \sum_{k=1}^{K} \sum_{i,j \in \Omega} |u_{i,j} - f^k_{i,j}| \right\}. \tag{4.3}$$

As the L^1 norm in the data term is known to be not optimal for Gaussian noise (we expect a small amount), we use the robust Huber norm [Huber, 1981] to estimate the error between sought solution and observations instead. The Huber norm is quadratic for small deviations (this is appropriate for handling normal distributed noise) and linear for larger errors, which corresponds to median-like behavior. The Huber norm is defined as

$$|t|_\epsilon = \begin{cases} \frac{t^2}{2\epsilon} & : \ 0 \leq t \leq \epsilon \\ t - \frac{\epsilon}{2} & : \ \epsilon < t \end{cases} . \tag{4.4}$$

Because of the range image driven alignment of the source information, undefined areas can be simply determined in advance for any geometrically transformed image f^k. Therefore, we support our formulation with a spatially varying term $w^k_{i,j} \in [0,1]^{WH}$, which encodes the inpainting domain. The choice $w^k_{i,j} = 0$ corresponds to pure inpainting at a pixel location (i,j). Note that in our case we only consider a binary case $\{0,1\}$, however, the strength of inpainting can be easily supported, *e.g.*, by incorporating the angle between an orthographic and real camera viewing direction. This context information would prefer objects, captured near to an orthographic view point, within the fusion step. Considering the encoded inpainting domain and the Huber norm, our modified energy minimization problem for redundant observations can now be formulated for the image domain Ω as

$$\min_{u \in X} \left\{ \|\nabla u\|_1 + \lambda \sum_{k=1}^{K} \sum_{i,j \in \Omega} w^k_{i,j} |u_{i,j} - f^k_{i,j}|_\epsilon \right\} . \tag{4.5}$$

In the following we highlight an iterative strategy based on a first-order primal-dual algorithm to minimize the non-smooth problem defined in (4.5).

4.3.3 Primal-Dual Formulation

Note that the minimization problem given in (4.5) poses a large-scale (the dimensionality directly depends on the number of image pixels, *e.g.*, for a small color image tile: 3×1600^2 pixels) and non-smooth optimization problem. Following recent trends in convex optimization [Nemirovski, 2004, Nesterov, 2005], we apply an optimal first-order primal-dual scheme [Chambolle and Pock, 2010, Esser et al., 2009] to minimize the energy. Thus we first need to convert the formulation defined in (4.5) into a classical convex-concave saddle-point problem. The general minimization problem is written as

$$\min_{x \in X} \max_{y \in Y} \{ \langle Ax, y \rangle \ + \ G(x) \ - \ F^*(y) \} \tag{4.6}$$

where A is a linear operator, G and F^* are convex functions and the term F^* denotes the convex conjugate of the function F. The finite-dimensional vector spaces X and Y provide a scalar product $\langle \cdot, \cdot \rangle$ and a norm $\| \cdot \| = \langle \cdot, \cdot \rangle^{\frac{1}{2}}$. By applying the Legendre-Fenchel transform to (4.5), we obtain an energy with the dual variables $p \in P$ and $q \in Q$ as follows

$$\min_{u} \max_{p,q} \left\{ \langle \nabla u, p \rangle - \delta_P(p) + \sum_{k=1}^{K} \left(\langle u - f^k, q^k \rangle - \delta_{Q^k}(q^k) - \frac{\epsilon}{2} \|q^k\|_2 \right) \right\}. \tag{4.7}$$

In our case, the convex sets Q and P are defined as follows

$$\begin{aligned} Q^k &= \left\{ q^k \in \mathbb{R}^{WH} : |q^k_{i,j}| \leq \lambda w^k_{i,j}, (i,j) \in \Omega \right\}, \; k = \{1 \dots K\}, & (4.8) \\ P &= \left\{ p \in \mathbb{R}^{WH} \times \mathbb{R}^{WH} : \|p\|_\infty \leq 1 \right\}, & (4.9) \end{aligned}$$

where the norm of the vector space P is defined as

$$\|p\|_\infty = \max_{i,j} |p_{i,j}|, \quad |p_{i,j}| = \sqrt{(p^1_{i,j})^2 + (p^2_{i,j})^2}. \tag{4.10}$$

Considering (4.7), we can first identify $F^* = \delta_P(p) + \sum_{k=1}^{K} \left(\delta_{Q^k}(q^k) + \frac{\epsilon}{2} \|q^k\|_2 \right)$. The functions δ_P and δ_{Q^k} are simple indicator functions of the convex sets P and Q defined as

$$\delta_P(p) = \begin{cases} 0 & \text{if} \quad p \in P \\ +\infty & \text{if} \quad p \notin P \end{cases} \qquad \delta_{Q^k}(q^k) = \begin{cases} 0 & \text{if} \quad q^k \in Q^k \\ +\infty & \text{if} \quad q^k \notin Q^k \end{cases}. \tag{4.11}$$

Since a closed form solution for the sum over multiple L^1 norms cannot be implemented efficiently, we additionally introduce a dualization of the data term with respect to G yielding an extended linear term with $\langle Ax, y \rangle = \langle \nabla u, p \rangle + \sum_{k=1}^{K} \langle u - f^k, q^k \rangle$. According to [Chambolle and Pock, 2010], the primal-dual algorithm can be summarized as follows: First, we set the primal and dual time steps with $\tau > 0$, $\sigma > 0$. Additionally, we construct the required structures with $u_0 \in \mathbb{R}^{WH}$, $\bar{u}_0 = u_0$, $p_0 \in P$ and $q^k_0 \in Q^k$. Following the solution presented in [Chambolle and Pock, 2010], the basic iterative scheme is then given by

$$\begin{cases} p_{n+1} &= \text{proj}_P \left(p_n + \sigma \nabla \bar{u}_n \right) \\ q^k_{n+1} &= \text{proj}_{Q^k} \left(\frac{q^k_n + \sigma(\bar{u}_n - f^k)}{1 + \sigma\epsilon} \right), \; k = \{1 \dots K\} \\ u_{n+1} &= u_n - \tau \left(-div \; p_{n+1} + \sum_{k=1}^{K} q^k_{n+1} \right) \\ \bar{u}_{n+1} &= 2u_{n+1} - u_n \; . \end{cases} \tag{4.12}$$

In order to iteratively compute the solution of (4.7) using the primal-dual scheme, pointwise Euclidean projections of the dual variables p and q onto the convex sets P and Q are required. The projection of the dual variable p is defined as

$$\text{proj}_P(\tilde{p}_{i,j}) = \frac{\tilde{p}_{i,j}}{\max(1, |\tilde{p}_{i,j}|)} \tag{4.13}$$

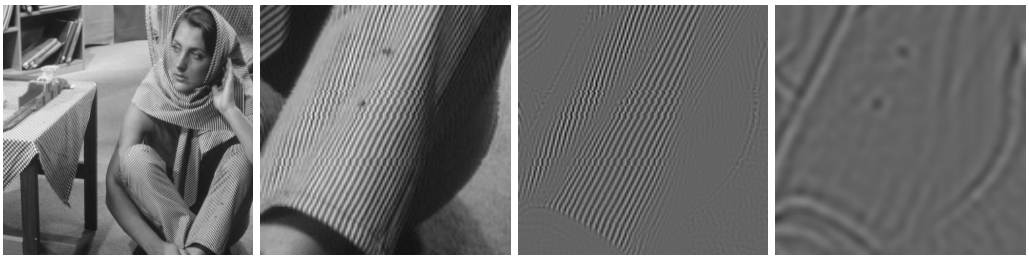


Figure 4.4: The result of decomposing an input image using the DTCWT. The magnitudes of computed filter coefficients are shown at two different levels of decomposition.

and the corresponding projections for the dual variables q^k with $k = \{1 \dots K\}$ are given by

$$\text{proj}_{Q^k}(\tilde{q}^k_{i,j}) = \frac{\tilde{q}^k_{i,j}}{\min(+\lambda w^k_{i,j}, \max(-\lambda w^k_{i,j}, |\tilde{q}^k_{i,j}|))}. \tag{4.14}$$

Note that the iterative minimization scheme mainly consists of simple point-wise operations, therefore it can be considerably accelerated by exploiting parallel hardware. The proofs for convergence and details can be found in [Chambolle and Pock, 2010].

4.3.4 Extension to a Wavelet-based Regularization

Since an orthographic image generation from color information with sampling distances of 8 to 15 cm requires an accurate recovery of fine details and complex textures, we extend our model by replacing the TV-based regularization (the simplest choice for structured sparsity) with a dual-tree complex wavelet transform (DTCWT) [Selesnick et al., 2005, Fadili et al., 2009]. The DTCWT is nearly invariant to rotation, which is important for a sophisticated regularization, but also invariant to translations and can be computed efficiently by using separable filter banks. The transform is based on analyzing the signal with two separate wavelet decompositions, where one provides the real-valued part and the other one yields the complex part. Figure 4.4 depicts the absolute values of the obtained filter coefficients at different levels of decomposition. Due to the redundancy in the proposed decomposition, the directionality can be improved, compared to standard discrete wavelets [Selesnick et al., 2005].

In order to include the wavelet based regularization into our generic formulation we replace the gradient operator ∇ by the linear transform $\Psi : X \to C$. The space $C \subset \mathbb{C}^D$ denotes the complex- and real-valued transform coefficients $c \in C$. The dimensionality of $\mathbb{C}^D$ directly depends on parameters, like the image dimensions, the number of levels and orientations. The adjoint operator of the transform Ψ, required for signal reconstruction,

is denoted as Ψ^* and is defined through the identity $\langle \Psi u, c \rangle_C = \langle u, \Psi^* c \rangle_X$.

Taking into account the TV-regularized energy minimization problem defined in 4.5, we can simply rewrite the functional for the image domain Ω as

$$\min_{u \in X} \left\{ \|\Psi u\|_1 + \lambda \sum_{k=1}^{K} \sum_{i,j \in \Omega} w_{i,j}^k |u_{i,j} - f_{i,j}^k|_\epsilon \right\}, \tag{4.15}$$

since the wavelet transform corresponds to a linear operation. According to the primal-dual scheme described in Section 4.3.3, we obtain a slightly modified iterative solution, where we now have to update the dual variable $c \in C$. Again, we construct the required structures with $u_0 \in \mathbb{R}^{WH}$, $\bar{u}_0 = u_0$, $c_0 \in C$ and $q_0^k \in Q^k$. The iterative scheme is defined as

$$\begin{cases} c_{n+1} &= \text{proj}_C \left(c_n + \sigma \Psi \bar{u}_n \right) \\ q_{n+1}^k &= \text{proj}_{Q^k} \left(\frac{q_n^k + \sigma(\bar{u}_n - f^k)}{1 + \sigma\epsilon} \right), \; k = \{1 \dots K\} \\ u_{n+1} &= u_n - \tau \left(\Psi^* c_{n+1} + \sum_{k=1}^{K} q_{n+1}^k \right) \\ \bar{u}_{n+1} &= 2u_{n+1} - u_n \; , \end{cases} \tag{4.16}$$

where the point-wise Euclidean projections of the dual variable c onto the convex sets C is given according to

$$\text{proj}_C(\tilde{c}_{i,j}) = \frac{\tilde{c}_{i,j}}{\max(1, |\tilde{c}_{i,j}|)} \; . \tag{4.17}$$

Note that an extension of the TV-L^1 formulation to vector-valued data, such as color information, is straightforward and requires only slight modifications in the implementation. So far, we have shown how multi-view information, such as color and height maps, sampled within a common view, can be fused efficiently by using energy minimization schemes. To round up a meaningful holistic description of the aerial imagery in the orthographic representation we finally focus on the fusion of redundant semantic classifications in order to accurately explain every gathered point on ground.

4.4 Fusion of Redundant Classification Information

As extensively described in Chapter 3, the first processing step of our semantic interpretation workflow involves a rapid pixel-wise explanation, performed on each image in the aerial dataset. In our case we use either a binary building classification, where we separate *building* structures from the *background*, or a full image explanation into the elementary object classes representing *building*, *water*, *tree*, *street* and *grass*. In addition, in the last

chapter we have shown that a combination of appearance and 3D cues is essential for a reliable classification at the level of images.

Since we are interested in a large-scale semantic interpretation of whole urban environments, we again fuse the intermediate results (now we have a probable class assignment) obtained for the original images within an orthographic view. Figure 4.3 depicts some redundant scene classifications projected to a common view point. Note that the dominant object class assignment is shown without a regularization procedure. It can be clearly seen that depending on the perspective camera view the mapped objects are classified with varying accuracy. Thus, we propose to fuse the raw confidences obtained from multiple view point in order to compute an improved interpretation result. In addition, a regularization may help to smooth the labeling with respect to the real object boundaries and to reduce the problem of wrongly assigned class labels. The used color coding for these object classes is given in Figure 3.13.

4.4.1 Accumulation of the Classifier Output

In order to estimate a class-specific probability for each pixel in the orthographic view, computed class distributions (the multi-class RF, constructed as a supervised classfier, provides class distributions for every observed feature instance) from different view points are accumulated and then normalized for the number of multi-view interpretations. Note that compared to standard classifiers such, *e.g.*, SVMs, where the outputs are distances to an estimated decision boundary, RFs inherently provide an efficient mapping of high-dimensional data to a probabilistic output (forming a normalized class distribution). Therefore, an aggregating of multi-view results is straightforward. Figure 4.5 depicts the pixel-wise accumulation result obtained by the proposed classification and fusion workflow. Due to classification and projection at the pixel level, the individual results shows a high granularity with respect to the dominant class confidences (first three rows), while the last row shows an improved semantic labeling. Nevertheless, the fusion step is performed at the level of pixels without considerations of important object boundaries and connected image regions. Thus, we introduce a regularization step to obtain a consistent final semantic labeling.

4.4.2 Refined Labeling Within the Orthographic View

Although our proposed feature representation considers some local context information, collected within a small spatial neighborhood, each pixel in the original images and especially in the fused case are explained almost independently. The problem of obtaining a smooth and spatially consistent labeling of the whole image scene can be again seen as the task of multi-class image segmentation (see Chapter 3), where each pixel value is selected

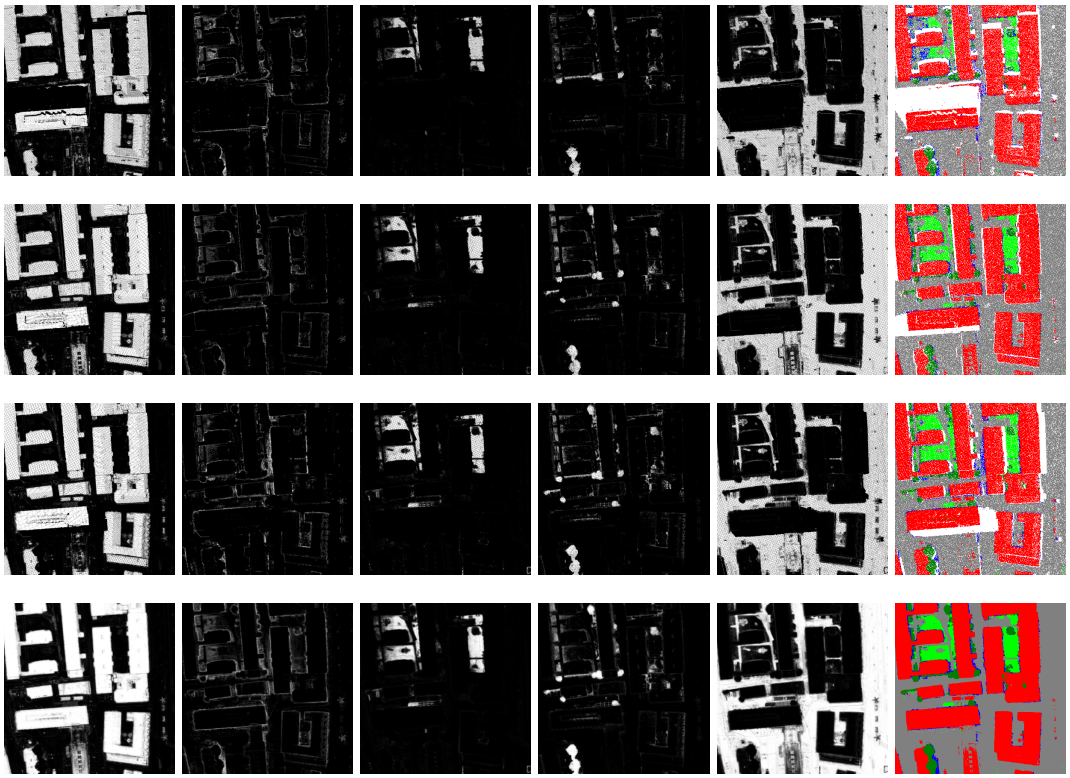


Figure 4.5: Semantic classification results projected to an orthographic view. The first three rows show the results obtained for redundant image tiles, while the last row depicts the aggregated confidences. The columns from left to right describe the confidences for *building*, *water*, *grass*, *tree* and *street*. The last column highlights a pixel-wise computation of the most dominant object class. Note that the pixel-wise fusion step considerably compensates for areas, where no interpretation is available (white pixels). In addition, the raw collection of multiple classification results significantly improves the final interpretation with respect to consistent assignment of object classes.

from a predefined pool of class labels. In our case the labeling procedure is supported by fused class distributions. In general, a segmentation problem into multiple object classes can be defined as a minimization of the Potts model [Potts, 1952], which was originally introduced to model phenomena of solid state physics. Recall that Ω defines an arbitrary image domain in $\mathbb{R}^{WH}$, then the continuous formulation of the Potts model for a partition into N distinct object classes can be written as

$$\begin{aligned} &\min_{s_i} \left\{ \lambda \sum_{i=1}^{N} \mathrm{Per}(s_i;\Omega) + \sum_{i=1}^{N} \int_{s_i} l_i dx \right\}, \\ &\text{s.t.} \bigcup_{i=1}^{N} s_i = \Omega, \quad s_i \cap s_j = \emptyset \; \forall i \neq j \; , \end{aligned} \tag{4.18}$$

The first term incorporates the length of the boundary of the segmentation s_i, while the second term considers the data at each point in the segment s_i. In turn, the scalar value λ defines the fidelity between data term and regularization (length of the boundary). The term $l_i \in \mathbb{R}^{WH}$ with $i = \{1 \dots N\}$ are the fused confidence maps for each object class and s_i is the resulting image partition for the object class i. In our case the confidence maps l_i are directly obtained by accumulation and normalization of corresponding multiple observations of the interpretation for different object classes.

In Chapter 3 we have largely applied minimization schemes, defined in discrete settings, to refine the semantic labeling. While discrete graph-based optimization techniques are in general hard to accelerate for a large amount of data, in this section we additionally suggest to use the continuous formulation of the Potts model. Recently, several solutions of the continuous formulation have been proposed [Chan and Vese, 2001, Pock et al., 2009, Olsson et al., 2009]. While Level-Sets [Chan and Vese, 2001] suffer from non-optimality, recent approaches are able to find globally optimal solutions of the Potts model. Olsson et at. [Olsson et al., 2009] proposed to reformulate the α-expansion moves [Boykov et al., 2001] in a continuous setting. Differently, Pock *et al.* [Pock et al., 2009] used a convex relaxation based on a primal-dual formulation of the TV functional. More importantly, they showed that their efficient primal-dual projected gradient algorithm enables a fast implementation on GPUs. Thus, in this section we additionally use the convex relaxation to minimize the energy defined in (4.18). According to [Pock et al., 2009], the minimization of the segmentation problem can be rewritten as a TV functional

$$\min_{u_i} \left\{ \lambda \sum_{i=1}^{N} \int_{\Omega} \sqrt{\nabla u_i(x)^T \mathrm{diag}(g(x)) \nabla u_i(x)} \; dx + \sum_{i=1}^{N} \int_{\Omega} u_i(x) l_i(x) \; dx \right\}, \tag{4.19}$$

where $u_i : \Omega \rightarrow \{0, 1\}$ is a binary labeling function with $u_i(x) = 1$, if $x \in s_i$ and $u_i(x) = 0$ else. The first term denotes the TV of the functions u_i and describes the corresponding anisotropic length of the segment s_i. The term $g \in \mathbb{R}^{WH}$ describes the edge penalty function, which enables a smooth labeling by taking into account strong edges, extracted from, *e.g.*, available color or height information (see Section 4.3). The approximated metric tensor $\mathrm{diag}(g(x))$ is used to incorporate the edge information into the energy functional. For each pixel the stopping function g is computed according to

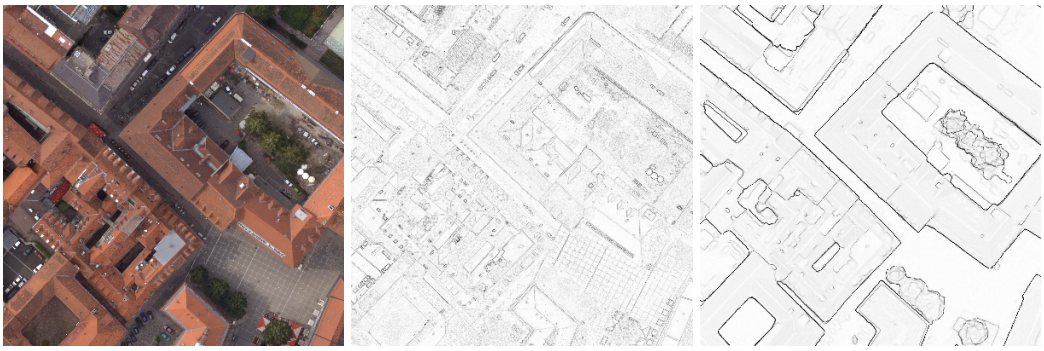


Figure 4.6: A computed penalty function for an arbitrary aerial scene. We derive the edge image either from color or height data. In particular for elevated objects the penalty terms, computed from height, will give an improved object delineation, since shadows and color changes may produce irrelevant edges.

$$g(x) = e^{-\alpha|\cdot|}. \tag{4.20}$$

The scalar α defines the smoothness and $\|\cdot\|$ is the pixel-wise magnitude, computed from simple color or height value distances. Figure 4.6 illustrates the stopping function, either resulting from color information or fused 2.5D height data. These functions are easily derived by using a standard Sobel mask filtering. The second term defines the data term, provided by the class confidence maps.

Since the space of binary functions u_i forms a non-convex set, the functional cannot be directly minimized using a convex optimization strategy, such as the efficient primal-dual algorithm. Pock *et al.* [Pock et al., 2009] proposed to relax the set of binary functions to a set of functions $u_i : \Omega \rightarrow [0, 1]$, which can take values between zero and one. They showed that this relaxation scheme yields globally optimal solutions for most practical problems. For the implementations details and proofs we refer to [Pock et al., 2009].

In the experiments we compare the results obtained for optimization schemes defined in the continuous and the discrete setting and show that the suggested refinement step ideally fits our holistic workflow and significantly improves the final semantic classification accuracy in terms of a qualitative and a quantitative evaluation.

4.5 Experiments

In this section we evaluate the described fusion approaches for color, height and semantic interpretation.

In a first step, the fusion models, based on energy minimization schemes taking into

account multiple scene observations for color and height information are investigated in a testing environment with artificially added noise and known ground truth data. The synthetic experiments focus on the image reconstruction capability in terms of the improved peak signal-to-noise-ratio (PSNR) by using an increasing number of input observations. The PSNR, computed for a ground truth image I_1 and a recovered or noisy image I_2, is defined as

$$PSNR = 10 \log \left(\frac{(\max I_1)^2}{1/|I_1| \sum (I_1 - I_2)^2} \right), \tag{4.21}$$

where the considered images are assumed to be normalized for the interval $[0, 1]$. Then, for both models we present real-world results obtained for aerial images, where we fuse multiple redundant image tiles showing urban environments. Since there exits no ground truth information, the evaluation is mainly performed by visual inspection.

Second, we show how the use of redundant semantic classifications influences the fused interpretation within an orthographic view. A final refinement step based on optimization schemes exploits both the corresponding color and height data to obtain a high-quality semantic interpretation, collected from multiple views.

4.5.1 Experiments on Height Data Fusion

In this section we first apply the described TV-L^1 fusion model (Equation (4.5)), capable to handle multiple input observations, to synthetically generated height data in order to evaluate the obtained results quantitatively. The synthetically constructed surface provides the ground truth information, that allows us to measure the deviation between the real height values and those obtained with our fusion model. In a second experiment we use our fusion model to fuse real-world aerial data, where we integrate redundant height field observations to form an improved representation of the mapped surface. Note that our holistic scene description also include a fused terrain model, however, the evaluation focuses on the surface model only.

Synthetic Experiments

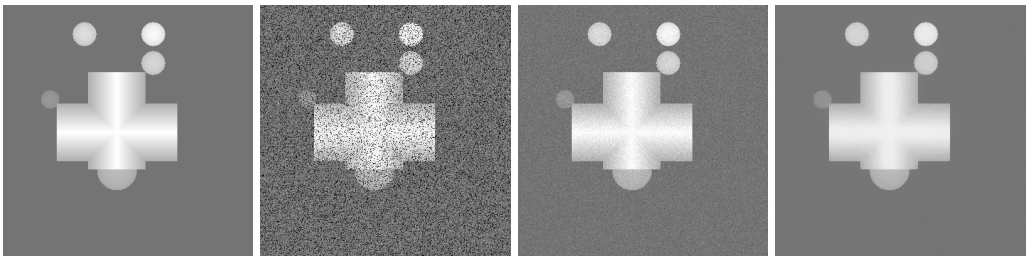
In order to evaluate the performance of the proposed model we selected an arbitrary scene from the web and constructed a synthetic test image consisting of an isolated building with a complex gabled rooftop and a some simplified tree-like structures. Our synthetic model is represented as an intensity image with a dimension of 400×400 pixels. The height values are normalized to the full range of 2^{16} available intensity values. Figure 4.7 shows our synthetically generated height model, the color information and the corresponding rendering.

Figure 4.7: A synthetically generated height model to evaluate the performance of our fusion model. From left to right: the height model, texture information and a rendering of the synthetic setup. We add artificial noise to the height field in order to produce multiple input observations. Since the ground truth is provided by the original model a quantitative evaluation can be performed in terms of the PSNR. We use a small amount of Gaussian noise, impulse noise and varying percentages of undefined pixel information.

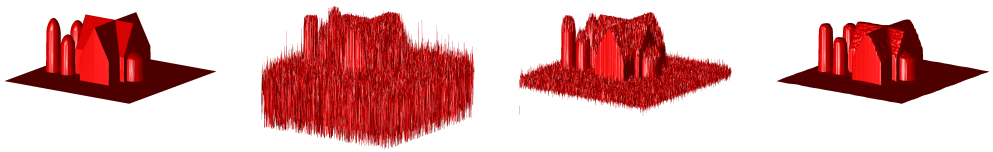
For the quantitative evaluation we add different types of artificial noise to the synthetic model: First, a small amount of Gaussian noise $\mathcal{N}(0, 0.01)$ and impulse noise (1%) is applied for all experiments. In addition, in order to evaluate the required inpainting capability of the fusion model defined in (4.5) a random set of a certain quantity of pixels is set to undefined. Thus, we run experiments with 10%, 25%, 50% missing pixel information. This simulates the real case for the aerial imagery, where undefined areas are caused by occlusions and moving objects. In order to investigate the influence of the redundancy the experiment is repeated for an increasing number of distorted input observations and a different quantities of undefined pixels. Table 4.1 summarizes the quantitative evaluation for 1, 5 and 10 input observations. In addition, we also calculate the PSNR for a mean and median computation, that do not provide a spatial regularization. From Table 4.1 we can observe that the result obtained with these simple methods improves with the number input observations. Nevertheless, the proposed fusion model significantly outperforms the mean and the median computation with more than 10 dB in all noise levels. The ratios are computed for the optimal parameter setup with $\epsilon = 1.0$, $\lambda = 1.0$, $\tau = 0.01$, and $\sigma = 1/8/\tau$, which has given the best result in a cross-validation framework. Some visual results are shown in Figure 4.8, where we use a number of 10 input observations. Although the tops of the mapped object cannot be reconstructed perfectly, the TV-regularization improves the quality of the fusion result with respect to the noise model. In addition, one can see that the result computed with TV-L^1 slightly suffers from stair-casing effects. To overcome these problems a Total Generalized Variation (TGV) regularization [Bredies et al., 2010] would further improve the quality of a height

	# Observations 1			# Observations 5			# Observations 10		
Undefined Pixels	10 %	25 %	50%	10%	25%	50%	10 %	25 %	50 %
Noisy image	14.1	10.7	8.0	14.1	10.7	8.0	14.1	10.7	8.0
Median	14.1	10.7	8.0	26.4	24.9	18.5	29.4	28.5	25.9
Mean	14.1	10.7	8.0	25.2	24.1	18.4	28.0	27.2	25.0
TV-L^1	29.0	27.1	21.2	36.8	36.0	34.1	**39.3**	**38.6**	**37.2**

Table 4.1: Quantitative evaluation of the height field fusion with multiple observations and a different amount of noise level. The evaluation is given in terms of the PSNR [dB]. The TV-L^1 model obtains the best noise suppression.



(a) Ground truth (b) 14.08 dB (c) 29.44 dB (d) 39.39 dB



(e) Ground truth (f) 14.08 dB (g) 29.44 dB (h) 39.39 dB

Figure 4.8: Some visual results computed from 10 distorted synthetic height observations. The first row shows the height model in the image space, while the second row depict the corresponding rendering. From left to right: the original height model, a noisy input observations, a result obtained with a median estimation and the fusion result of the TV-L^1 model. One can see that both the median and TV-L^1 model successfully reduces the outliers. However, we can notice that a regularization is essential for a reliable fusion of multiple height maps.

map fusion since the model better supports piecewise affine image structures.

Experiments on Real Height Data

Next, we apply our TV-L^1 model to real-world height maps extracted from the large-scale aerial imagery. Since no ground truth data is available for the three datasets *Dallas*, *Graz* and *San Francisco* we mainly concentrate on a visual inspection of the computed fusion results. Note that an quantitative evaluation could be performed by, *e.g.*, comparing the result to a point cloud taken with LiDAR. In Figure 4.9 computed image tiles for *Dallas* are depicted. In order to handle the large amount of available data we divide the fusion into 1600×1600 pixels image patches. The $2.5D$ fusion result is computed from an average of 10 input observations. One can see that although the proposed model works fine for the majority of the challenging building structures, specular facades and insufficient overlap between the aerial images, causes massive artifacts (bottom of the third image), which cannot be recovered completely. Although an increased strength of regularization would account for such massive outliers, the structures of the remaining objects then get over-smoothed. The real-world height models given in this thesis are computed for $\lambda = 1.0$.

Figure 4.10 illustrates the fusion result obtained for a single scene of *Graz*. The images provide a pixel size of approximately 8 cm. From the rendering one can see that fine details, such as dormers, are nicely preserved, while the major rooftop structures are correctly delineated. Compared to the *Dallas* scene, the isolated high-rise building is captured entirely since there are no neighboring building structures, which annoy the correspondence computation in the dense matching process.

A computed strip for *San Francisco*, consisting of 10×3 individually processed image patches, is shown in Figure 4.11. Although *San Francisco* is embedded in a hilly terrain the surface model can be reconstructed nicely. A tight fusion of a single image patch, representing color and height information, requires less than 2 minutes on a standard graphics card. Note that our large-scale workflow can be applied to process any quantity of input images due to the block-processing in smaller image tiles.

4.5.2 Experiments on Color Fusion

The next step of our holistic interpretation workflow aims at fusing the available color information into an orthographic view. We therefore apply our convex fusion model, including the wavelet-based regularization, to generate high-quality real-world ortho-images. In this section we again apply our models (Equation (4.5) and (4.15)) to synthetic data, then we use it for integrating redundant color observations. To show the benefit of the wavelet-based regularization, we compare the obtained results to those of the TV-L^1 model, the non-regularized mean and median fusion.

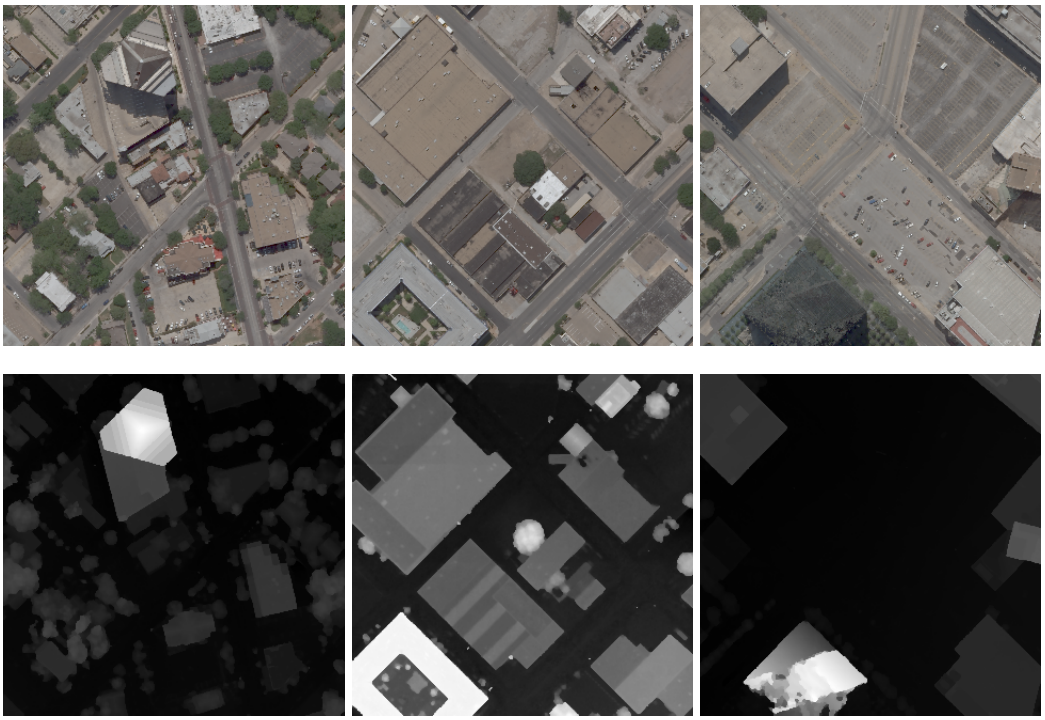


Figure 4.9: Some orthographic color and height fusion results for *Dallas*. The first row shows the fused color information computed with the wavelet-based approach presented in 4.3. The second row depicts the corresponding height fields with a pixel size of approximately 15 cm.

Synthetic Experiments

In order to evaluate the quality of our joint inpainting and denoising model, we first apply our formulation to the task of color image recovery using a single input observation only, where parts of the test images are replaced by black (undefined) pixels. To simulate the occlusions in the aerial imagery, caused by the transformation of perspective images to an orthographic view, we thus artificially set a certain amount of pixel to undefined by using randomly generated line structures. Figure 4.12 depicts an image distorted with 20 random lines, and the results obtained for recovery using the TV-L^1 and the DTCWT-regularized model. Although the structures are not recovered completely, the TV-norm in the smoothness term yields cartoon-like results compared to the improved wavelet-based regularization. The results for a quantitative evaluation are summarized in Table 4.2. Due to the randomness of this experiment we compute the statistics for the PSNR over 10 runs. One can see that the model with the wavelet-based regularization performs best for different numbers of random lines.

To show the performance with respect to recovered fine details, our second experiment

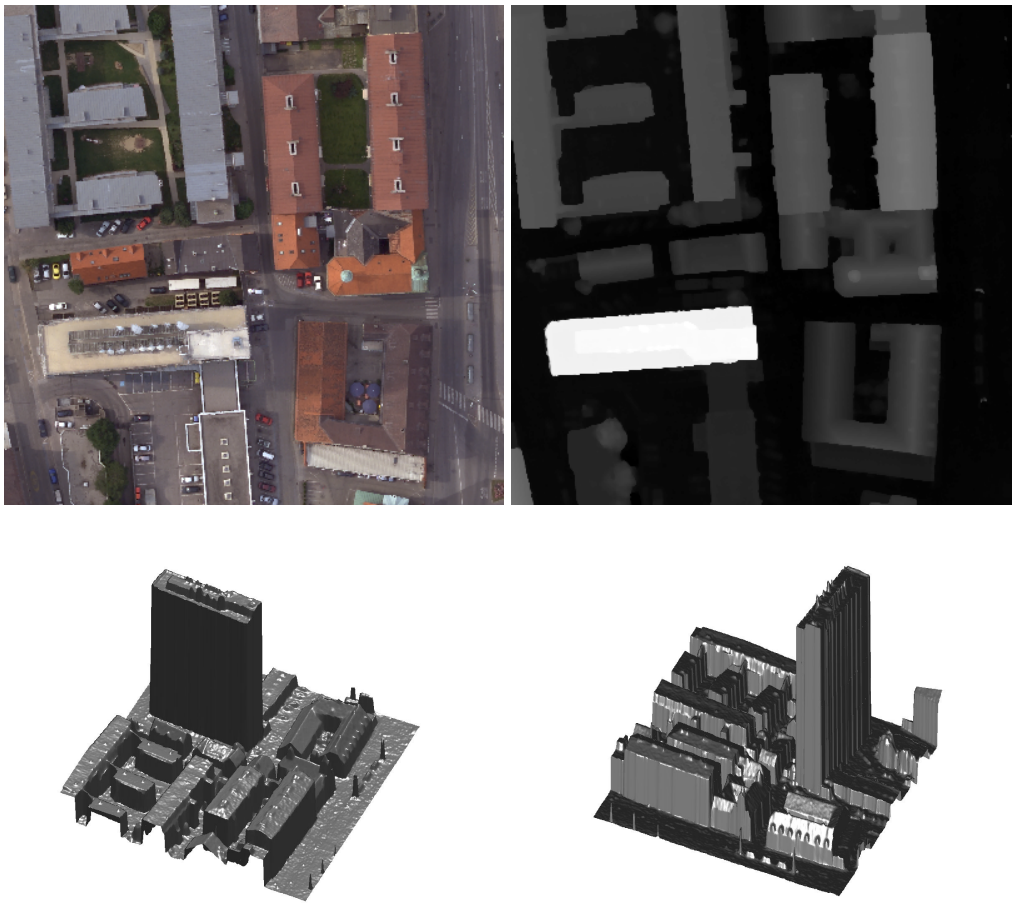


Figure 4.10: An orthographic view of a *Graz* scene. The first row shows the color and height fusion result for a 8 cm pixel size, while the second row depicts a rendering of the fused height field. Note that fine details, like structures on the rooftop, are preserved accurately.

Number of lines	Noisy Image [dB]	TV-L^1 [dB]	Wavelet [dB]
10	17.32 ± 0.98	31.20 ± 0.50	32.08 ± 0.41
20	14.35 ± 0.75	30.26 ± 0.65	31.26 ± 0.57
40	12.04 ± 0.41	28.82 ± 0.69	29.81 ± 0.41

Table 4.2: Quantitative evaluation of the image recovery from a single input observation. We average the PSNR over 10 runs and compute the ratios for different numbers of random lines. The wavelet-based regularization obtains the best noise suppression.

investigates the denoising and inpainting capability of our proposed model using multiple images with synthetically added noise. Here, we take the Barbara and the Lenna gray-

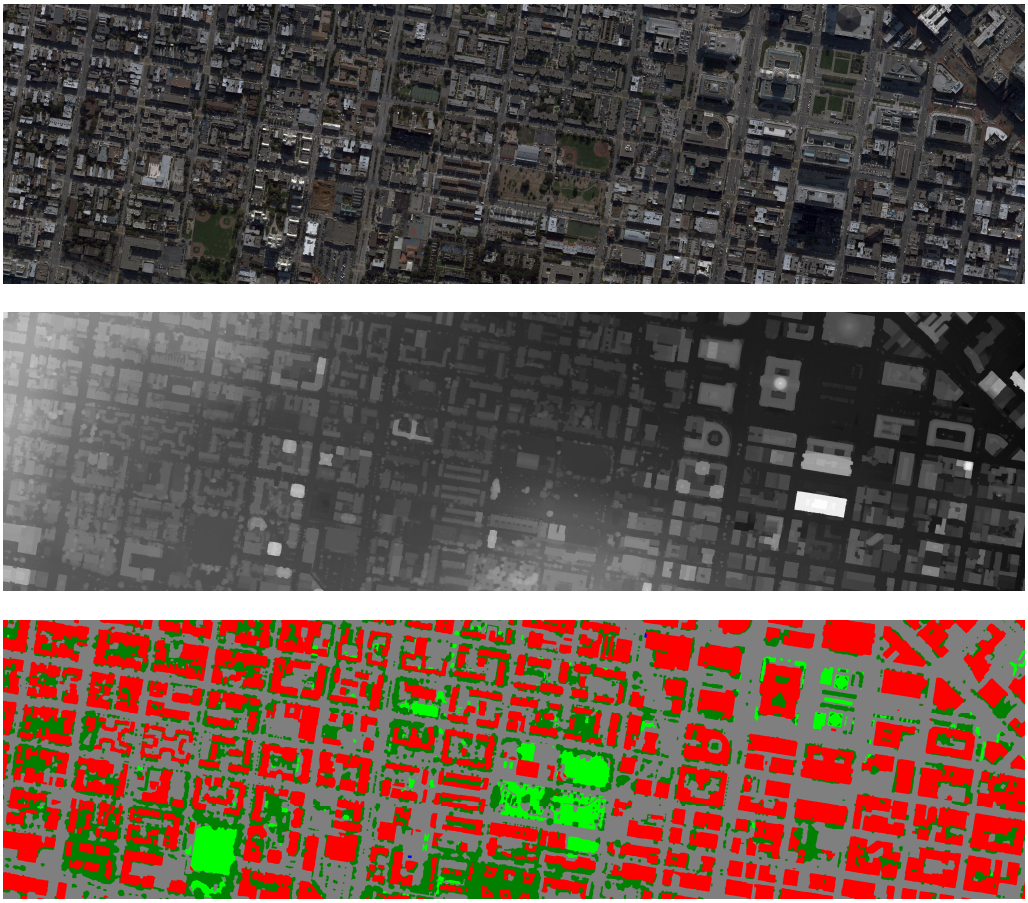


Figure 4.11: A computed strip of *San Francisco* with a GSD of 15 cm. A number of 10×3 patches are combined to from a holistic description of the mapped area. The depicted image covers an area of approximately 2500×700 meters. The interpretation results from a refinement procedure using the continuous formulation of the Potts model [Pock et al., 2009].

valued images. These 512×512 pixels test images contain fine structures and highly textured areas. In order to imitate the expected noise model, we add a small amount of Gaussian noise $\mathcal{N}(0, 0.01)$ and again replace a specified percentage of pixels with undefined areas (we use 10%, 25% and 50%). In Figure 4.13 distorted Barbara images are shown for a different amount of undefined pixels.

An evaluation in terms of the PSNR for different amounts of undefined pixels and quantities of input observations is summarized in Figure 4.14. We compare our model to the TV-L^1 formulation, the mean and the median computation. For the TV-L^1 and our model we present the plots computed for the optimal parameters determined by an exhaustive search over meaningful parameter settings. One can see that our joint inpainting

Figure 4.12: Image recovery from a single input observation, which is distorted with random lines. Form left to right: the noisy image, the inpainting result computed with the TV-L^1 model and the result obtained with the wavelet-based model.

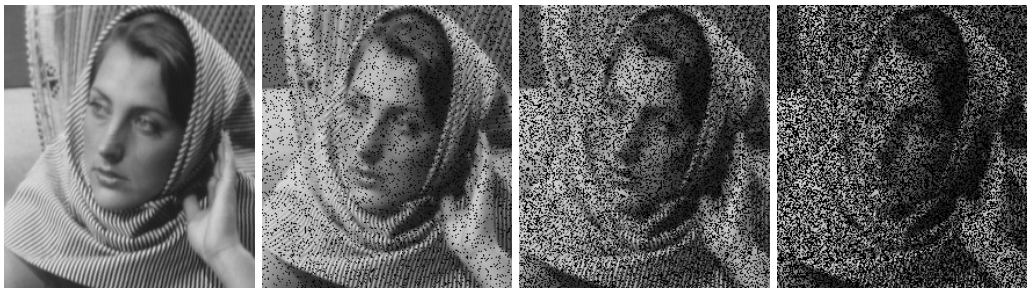


Figure 4.13: Gray-valued Barbara images with artificially augmented noise. We use a small amount of Gaussian noise $\mathcal{N}(0, 0.01)$ and different percentages of undefined pixel information (10%, 25%, 50%).

and denoising model, using the parameter setting $\tau = 0.05$, $\sigma = 1/8/\tau$, $\epsilon = 0.1$, $\lambda = 0.8$ (Lenna), $\lambda = 1.2$ (Barbara) and 3 levels of wavelet decomposition (we use $(13, 19)$-tap and Q-shift 14-tap filter kernels, respectively), performs best in all noise settings. Moreover, it is obvious that an increasing number of input observations significantly improves the result. Compared to the TV-L^1 model, the wavelet-based regularization improves the PSNR by an averaged value of 2 dB. In addition, we can observe that the plot, computed for the Lenna image, improves slightly faster with the number of input observations since this image contains less high-frequency textured areas than the Barbara image.

Fusion of Real Color Images

Our next experiment focuses on the fusion of aerial color images in order to produce high-quality ortho-photos. Due to resource limitations in the processing, we collect the huge amount of provided information within smaller image patches (1600×1600 pixels).

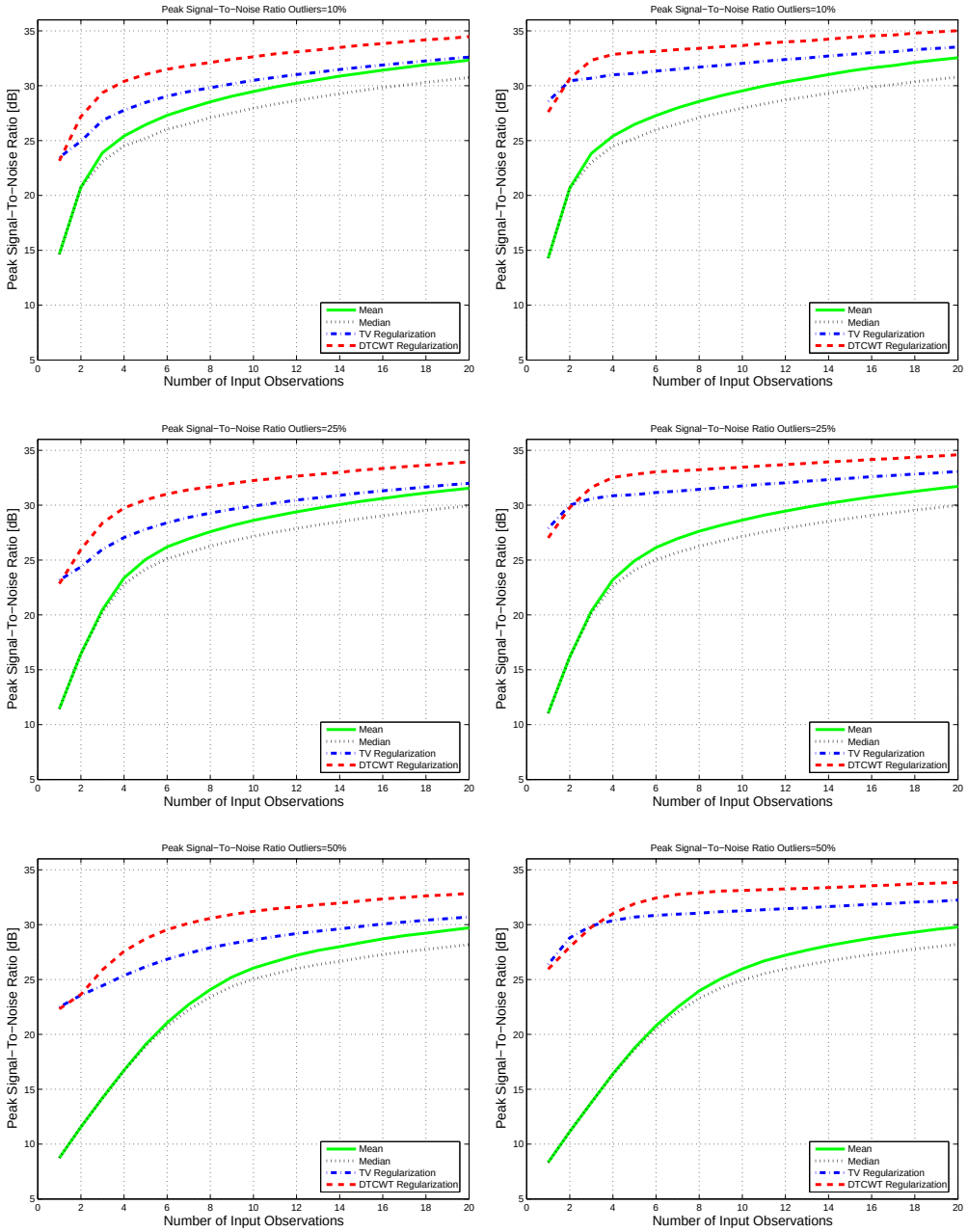


Figure 4.14: Quantitative results for the Barbara and the Lenna image: the PSNR depending on synthetically added noise and a varying number of input observations. The first column depicts the plots obtained for the Barbara image, while the second column the results for Lenna. The proposed model using the DTCWT-based regularization yields the best noise suppression for both test images.

As shown in Figure 4.3, the high overlap in the aerial project yields redundant scene observations, represented in the RGB color space. Depending on the provided overlap in the aerial imagery, *e.g.*, 60% and 80%, the mapped area provides up to 15 redundant observations for the same scene.

Figure 4.15 depicts four redundant details for *Graz*, where large regions are missing (black pixels) due to occlusions or sampling effects. One can see that both the TV-L^1 and the wavelet-based model fill these areas successfully by exploiting the redundant color information. However, the wavelet-based regularization preserves fine image structures with an improved natural appearance.

Figure 4.16 shows the fusion result generated for three observations of a low-rise building embedded in a suburban environment. The ortho-projected images contain only a small amount of undefined pixels mainly due to low-height structures and sufficient coverage of the perspective images. Both models yield suitable fusion results, however, in vegetation-occupied regions the result appears over-smoothed. This is mainly caused by small 3D reconstruction in these areas caused by natural movements of the vegetation, *e.g.*, caused by the wind. To overcome the problem of smoothed areas, an integrated combination of semantic interpretation and image fusion would give improved results. Figure 4.17 shows a generated ortho-photo for a region with a size of 128×128 meters. In Figure 4.18 and 4.19 large-scale results are given. Note that a computation within smaller image patches enables a quick computation of the ortho-photos with multiple cores.

Removal of Non-Stationary Objects

So far, we have shown that redundant color information can be successfully fused to a single image without undefined areas. However, non-stationary objects, such as moving cars, disturb in orthographic image generation. Thus, we additionally use our fusion model to remove cars by simultaneous inpainting. Car detection masks for aerial images can be obtained efficiently by various approaches to choice, *e.g.*, [Viola and Jones, 2004, Grabner et al., 2008, Mauthner et al., 2010]. Figure 4.32 shows some car detections results obtained for scene of *Graz*. In a first step we apply a trained car classifier (we use an efficient approach described in [Mauthner et al., 2010]) to some overlapping perspective intensity images. Then, the acquired confidence maps (we treat the car detection as a binary detection task), together with the color information, are transformed to a common view by using the available range and camera data (as described in Section 4.2). A combination by simply aggregating the confidences and thresholding for a suitable activation threshold yields the final car detection mask. Then, having the mask defines an extended inpainting domain within the fusion model. It is obvious that the detected cars produce large regions of undefined pixels, where even the high redundancy of the images do not

Figure 4.15: Obtained fusion results for a detail of *Graz*. The first rows shows redundant scene observations with many undefined pixels, mainly caused by occlusions and sampling effect. Both the TV-L^1 (left, second row) and the wavelet-based fusion model (right, second row) obtain accurate results, however the image-priors provided by the wavelet decomposition shows an improved appearance.

provide useful color information. Hence, we estimate an individual color prior for each car detection or a group of recognized cars, by computing the median over color samples collected along individual computed car masks. Besides that moving cars are inherently removed due to the high redundancy (see Figure 4.20), parking vehicles can be successfully removed from the ortho-images by using the estimated, however constant, color value priors. In order to obtain an improved filling of the detected car areas we outline an improved strategy, that is mainly inspired by the work of Hays and Efros [Hays and Efros, 2007].

The main idea is based on performing scene completion with respect to the computed detection mask by using a pool of potential exemplars. Thus, we randomly collect image patches (the dimension is adapted for a common car length) and apply basic image transformations, including rotation and translation, in order to synthetically increase the pool of candidates. To find the best matching candidate for each detected car we compute

Figure 4.16: A fusion results for a country residence, located in the surroundings of *Graz*. The first rows shows three scene observations with a relatively low number of undefined pixels. Although the fusion works fine for both models, the vegetation regions appear over-smoothed.

a sum of weighted color distances between a masked detection and each exemplar. The weighting additionally prefers pixel locations near the mask boundary and is mainly derived by using a quick distance transform [Felzenszwalb and Huttenlocher, 2004a]. The detection mask with overlaid exemplars is then used as an additional input observation within the color fusion model. Obtained removal results are shown in Figure 4.31. Note that this approach works only for isolated car detection results and if potential inpainting candidates are possible to extract.

4.5.3 Experiments on Classification Fusion

In Chapter 3 we have extensively investigate how the tight use of appearance and 3D height information can be utilized to compute an initial semantic interpretation of aerial

Figure 4.17: A wavelet-based fusion result for a larger region (128×128 meters). We integrated a total number of 8 redundant scene observation. Note that the fusion step successfully removes moving objects, such as cars.

images. The experimental evaluation showed that the combination of both is essential for an image classification. However, the semantic explanation is treated separately for each image in the aerial dataset. Figure 4.22 shows computed classification results into the five object classes for some overlapping aerial images, which can be also interpreted as a multi-view large-scale semantic interpretation since each single image provides a number of 11500×7500 interpreted pixels. Note that due to relevance we only consider the interpretation results obtained for a tight combination of appearance and elevation measurements.

In this section, we mainly focus on demonstrating the benefit of exploiting the redundancy for the generation of a final semantic interpretation within an orthographic view. Similar to the last chapter, we consider an interpretation into the two object classes (*building* and *background*) and into the five classes, where the image content is separated into *building*, *water*, *tree*, *grass* and *street*.

We first investigate how the fusion from multiple interpretation results influences the final classification in terms of correctly classified pixels. Then, we compare different optimization methods to refine the results by taking into account derived object boundaries. Finally, we compare obtain classification accuracies to the ones obtained with an

Figure 4.18: Some wavelet-based fusion results obtained for the inner city of *Graz*. We compute the fusion for image patches with a size of 1600×1600 pixels. Every pixel in the image patches represents a point on ground with a size of approximately 8×8 cm.

Figure 4.19: Some wavelet-based fusion result computed for a small part of *San Francisco*, where every pixel describes a point on ground with a size of 15 cm.

existing land-use classification workflow [Gruber-Geymayer et al., 2005]. Figure 4.21 summarizes the intermediate results of our fusion and refinement procedure. The first row depicts several corresponding semantic interpretation results, that are initially transformed to a common view. Then, a pixel-wise accumulation of the class probabilities yields the input for the refinement step. The results, before and after the refinement, are shown in the second row. Additionally, we exploit color or height information to enable a consistent final class labeling with respect to real object boundaries.

In order to perform a quantitative evaluation of our classification and fusion strategy

Figure 4.20: Car inpainting by using constant color priors. From left to right: The original perspective image, the fusion result without using a car detection mask and the inpainting result obtained and for the wavelet-based model.

to obtain a final interpretation, we have to additionally generate ground truth information for an orthographic view. Due to a balanced existence of the five considered object classes 10×6 image tiles (each tile covers an area of 128×128 meters) are largely labeled by different operators to provide ground truth data at the pixel level. However, bordering and transition regions are not assigned with one of the object classes. Clearly an interactive labeling scheme, as the strategy described in [Santner et al., 2010], would account for an improved quantitative evaluation of our proposed approach. For the remaining aerial projects *Dallas* and *San Francisco*, sets of 6×6 and 6×5 ground truth maps are generated by a rigorous procedure of manual labeling, respectively. In Figure 4.21, the labeled ground truth image is shown for the *Kunsthaus* scene.

The Fusion of Redundant Classifications

The first experiment investigates how redundancy influences the interpretation accuracy in terms of undefined pixels and correctly classified pixels. In order to determine the classification accuracy in the orthographic view, we extract a subset of labeled ground truth maps (25 image patches for the *Graz* dataset), covering an area of approximately 500×500 meters. Due to a varying number of occluded pixels, we repeat the fusion step 20 times (shown as blue dots) with random sets of selected observations. It is obvious that the percentage of undefined pixel significantly decreases with the number of involved observations. Note that the remaining undefined pixels could be entirely eliminated by an additional refinement step. This experiment also shows that the challenging task of distinguishing between *water* and *street* regions, covered with shadows, benefits from aggregating multiple redundant classifications. In addition, the classification rate of small *grass*-covered areas can be improved by using multi-view observations.

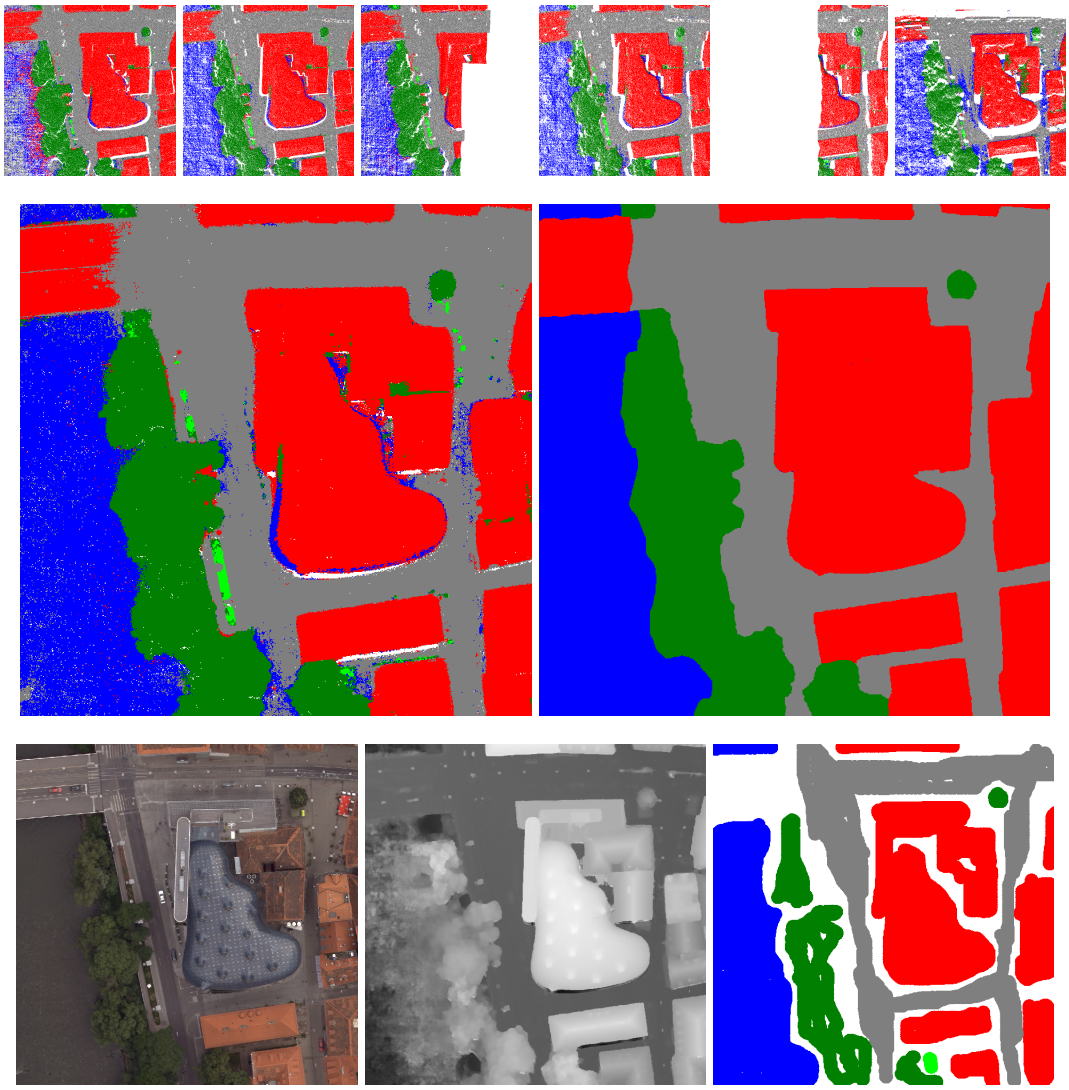


Figure 4.21: A semantic interpretation result for the *Kunsthaus* scene: The first row shows redundant semantic classifications projected to a common orthographic view. Note that these images include many undefined regions (white pixels) caused by occlusions and varying viewpoints. The raw accumulation over redundant image patches and the final refined classification result are given in the second row. The third row shows a corresponding fused color image, 2.5D height information and the hand-labeled ground truth map.

Refinement of the Semantic Interpretation

For practical applications it is essential to obtain an accurate semantic classification results with respect to real object boundaries. We therefore compare the results obtained for the

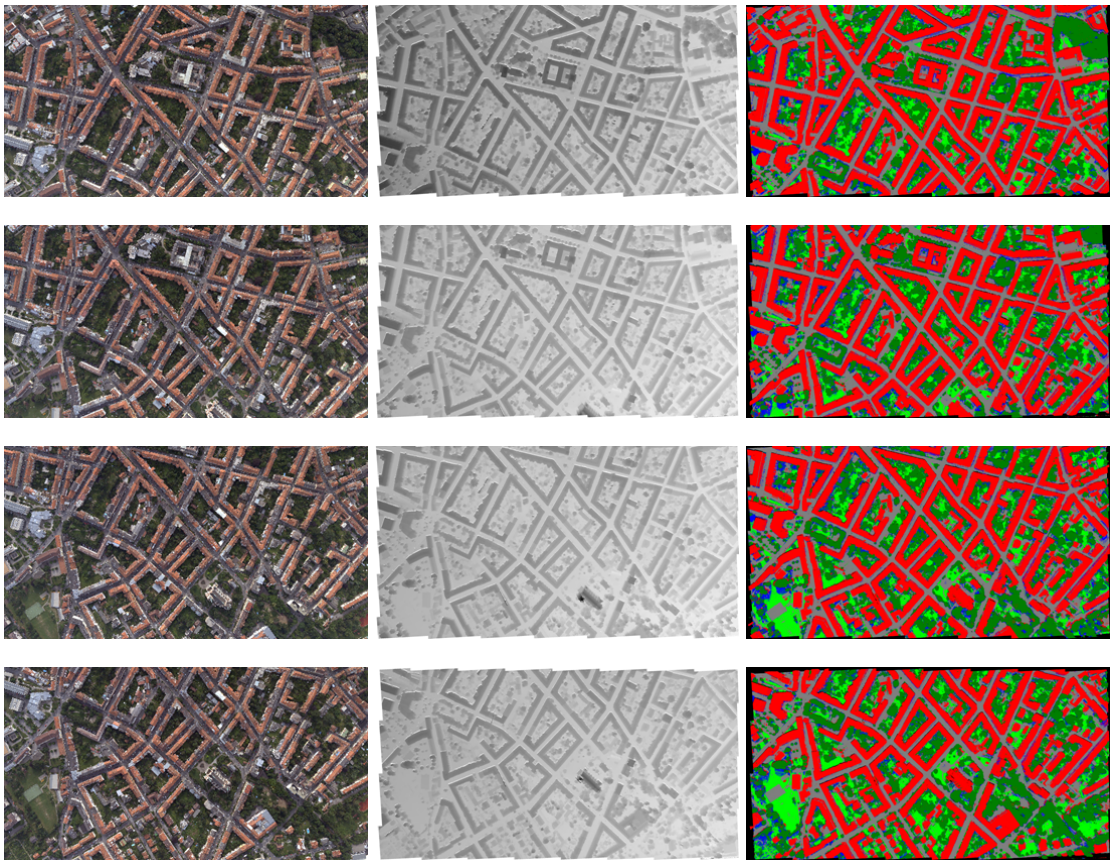


Figure 4.22: Overlapping aerial images: We exploit color (first column) and dense matching results (second column) to derive a highly redundant semantic classification (third column). Each image has a dimension of 11500×7500 pixels and was processed in a few minutes.

differently introduced refinement methods. In particular, we compare the results of a discrete optimization strategy defined on a four-neighborhood connected graph, a refinement procedure performed on an adjacency graph by incorporating the available super-pixel segmentation and the multi-class segmentation approach highlighted in Section 4.4, that is formulated in the continuous domain. For both the four-connected neighborhood and the super-pixel graph we minimize the energy functional by using the well-established α-expansion [Boykov et al., 2001]. Note that these methods are computed on a conventional single-core machines, since a parallel computation is hard to achieve. In contrast, the continuous formulation of the Potts model (4.19) can be minimized efficiently by exploiting, *e.g.*, multiple cores of a GPU. For the binary labeling case - the building classification - the refinement strategies yield globally-optimal solutions, while in the multi-class problem only an approximation is achieved in a worst-case scenario.

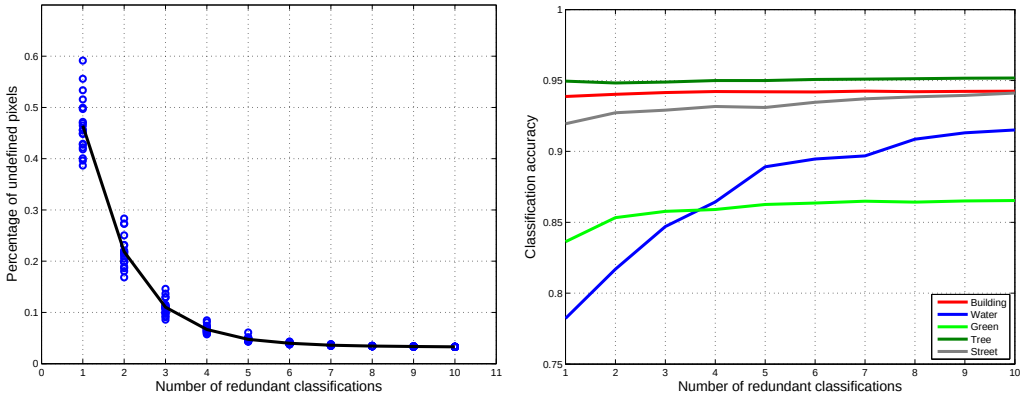


Figure 4.23: Fusion of redundant semantic classifications: The first plot shows the percentage of undefined pixels as a function of involved observations. The second plot depicts the obtained accuracy in terms of correctly classified pixels for an increasing quantity of involved semantic classifications. Due to varying occlusion, we average the rates over 20 runs. It can be clearly seen that the classes *water* and *grass* benefit from a fusion of redundant classifications. Due to high reconstruction errors (mainly composed of undefined depth values) in strongly moving *water* areas the interpretation workflow yields uncertain or missing class decisions (see *e.g.*the amount of white (undefined) pixels in the first row of Figure 4.21). Hence, the integration over multiple classification results in an improved interpretation, and therefore in an increased accuracy.

Table 4.3 compares the computed classification rates for the three datasets. Note that we skip the evaluation of the water class for the aerial dataset *Dallas* and *San Francisco* since there are no representatively large enough water areas available in the test region. Figure 4.27 represents these results as confusion matrices. It is perfectly clear that also a single interpretation, performed on fused color and height information could be used instead of using redundant classification. While such an approach would only provide an interpretation of structures that are clearly visible in the ortho-view, the proposed workflow also offers an initial classification of vertical objects such as house walls. This information could be utilized to derive accurate GIS information or to determine the exact location of the brickwork.

Figure 4.24 illustrates the refinement result obtained for a *San Francisco* scene by using different optimization strategies. In the case of the multi-class interpretation the edge information, required for the object delineation, is directly derived from the color information. Considering (4.20), the smoothing parameter is set to $\alpha = 1.0$ for all experiments performed for the five-class problem. The trade-off between data fidelity and regularization is set to $\lambda = 10.0$ for both discrete optimization schemes, while we used $\lambda = 0.5$ for the minimization of the continuous formulated functional. Note that feeding

the CRFs with the negative logarithms of the class probabilities has given an improved interpretation result in terms of correctly labeled image regions.

Taking into account the quantitative evaluation and the visual inspection of the obtained results for the fusion of the semantic interpretation, we can conclude that the continuous formulation of the Potts model yields a reasonable refined labeling with nearly no parameters to adjust within an attractive computation time of approximately 20 seconds per image tile with a size of 1600×1600 pixels. Compared to the method of Boykov *et al.* [Boykov et al., 2001] (16 seconds), the refined labeling on a super-pixel neighborhoods can be performed quickly (3 seconds), however the super-pixel generation itself introduces a significant loss of computation time. In addition, from a practical point of view a time-consuming parameter tuning is required for both discrete methods in order to achieve comparable refinement results. A full semantic interpretation obtained for a strip of *Dallas* is shown in Figure 4.26. The applied continuous formulation of the Potts model largely preserves objects boundaries, like building edges, by using a color-driven stopping function. Figure 4.21 summarizes the different steps of our refinement procedure.

Some observed special cases are illustrated in Figure 4.25. One can see that even the final labeling cannot entirely compensate large classification errors. These errors are mainly caused by wrongly generated elevation measurements with coincidentally similar appearance to the *street* class. Obscured image regions are sometimes classified as *water* since the wavelength of mapped water regions is very similar to the ones of shadows. Although a higher radiometric resolution would account for these problems, the available image timestamps together with the known 3D scene geometry could be utilized to predict shadowed areas and to improve the final interpretation result. Furthermore, from Figure 4.25 we can observe that the tight integration of appearance and 3D height enables a correct assignment of the *building* class, even if the rooftops are covered with grass, provided that these special cases are trained with the classifier.

For the binary building classification task, averaged final rates of more than 90.0% are obtained for *Dallas* (*building*: 97.9% and *background*: 96.2%), *Graz* (94.7% and 95.2%) and *San Francisco* (91.6% and 96.5%). These rates are computed by comparing the refined results (we use the energy minimization scheme defined in the continuous domain (4.18)) to the ground truth labeling. Compared to the five-class problem, where we distinguish between *building*, *water*, *grass*, *tree* and *street*, one can observe that the binary classification yields slightly increased rates of the *building* class. However, for the two-class case, mainly an elevated object class, namely the *building* class, is discriminated from *background*. This allows us to exploit edge information, derived from the fused 2.5D height data, as penalty function. Such a penalty function, applied to the five-class interpretation, would destroy, *e.g.*, the transitions between grass and street regions since there is no appreciable deviation in the computed elevation measurements. The effect

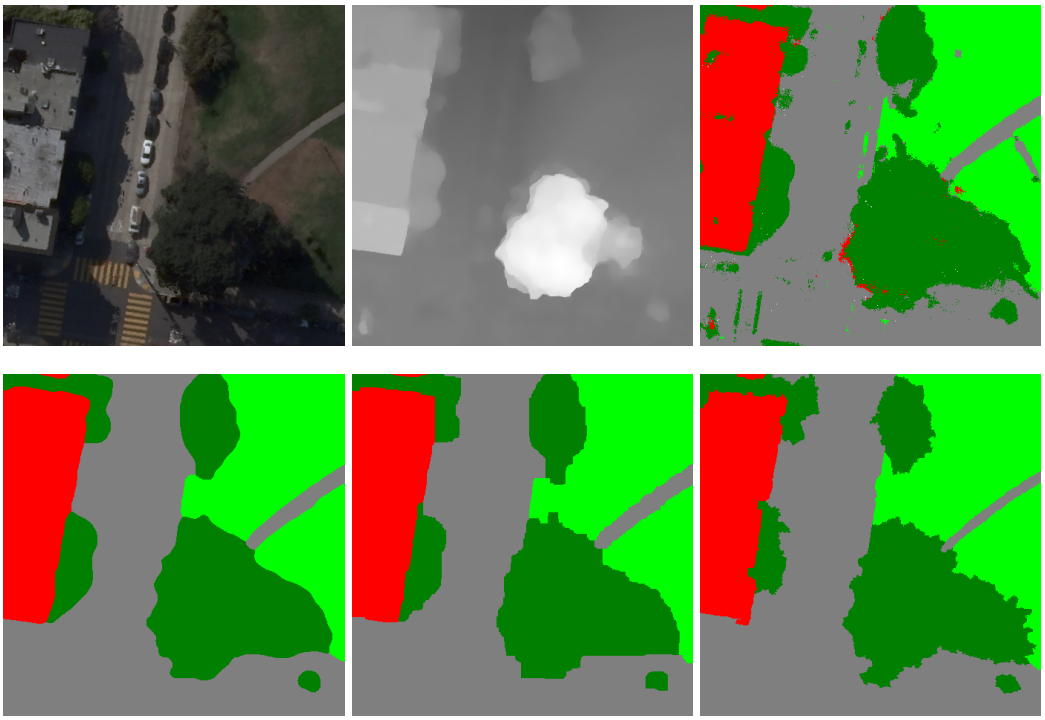


Figure 4.24: A visual comparison of various refinement strategies. The first row depict the color image, the corresponding height field and the raw classification result without a refinement step. The refinement results are given for the continuous version of the Potts model, the discrete optimization formulation using a four-connected graph and the optimization on a super-pixel neighborhood graph.

of using color- and height-driven stopping function within the continuous Potts model is shown in Figure 4.28. Due to different value ranges of the height data the smoothing parameter is set to $\alpha = 20.0$. It can be clearly seen that the height-driven refinement step obtains improved results concerning the real building edges. In particular, shadowed areas benefit from the regularization with a height-driven penalty function.

Comparison to an Existing Workflow

Finally, we compare the results produced with the proposed workflow to the available land-use classification for *Graz*, computed with the method described in [Gruber-Geymayer et al., 2005]. This method is mainly based on three processing steps: Without consideration of the 3D height information, in a first step the raw aerial images are coarsely divided into initial object classes, like solid, shadow, water and vegetation, by using SVM classifiers, that are trained in a supervised manner. Once the corresponding range data is computed from the overlapping perspective images, these initial land-use classifications

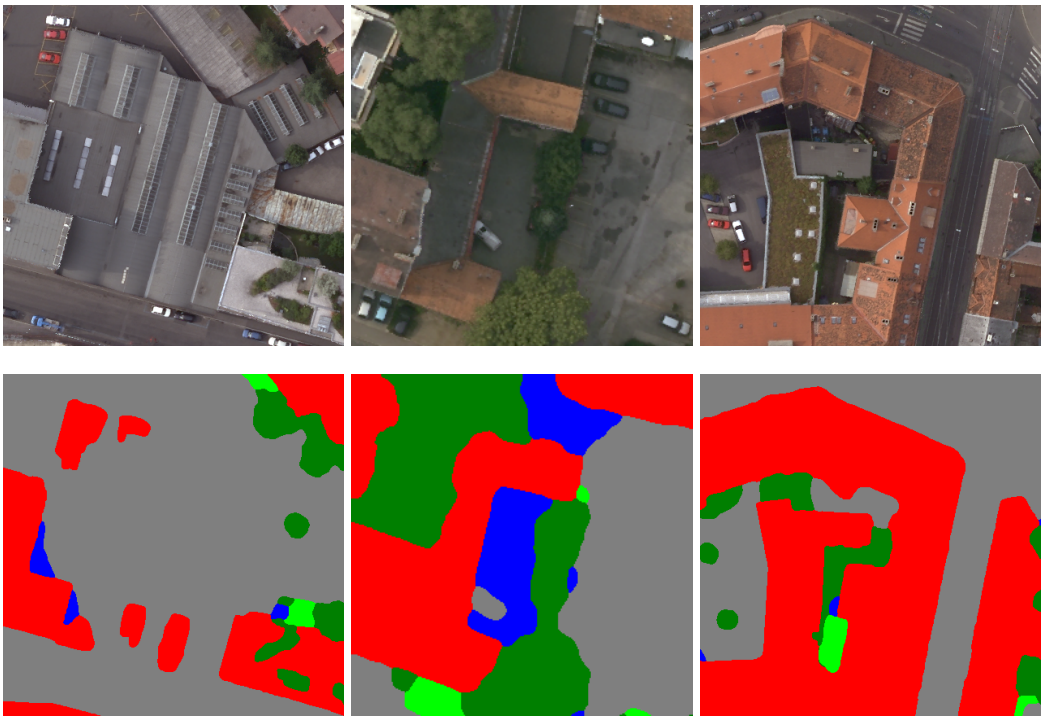


Figure 4.25: Some special (failure) cases. Even the refinement step cannot compensate large classification errors, caused by an erroneous DTM extraction and similar spectral wavelength. The spectral range of shadowed regions is very similar to the ones of the water class. The rightmost sample shows that even grassed rooftops can be assigned a correct class label.

are aligned within an orthographic view by means of the camera data and provided depth values. As a second step, static prior information is used to separate the initial vegetation classification into *grass* and *tree* or to split the solid areas into *street* and *building*. Third, and similar to our approach, a multi-class segmentation based on an optimization scheme is applied in order to refine the final class labeling.

In order to compare the performance of the very different workflows, we first align both results into a common world coordinate system with a corresponding GSD at 8 cm. Figure 4.29 shows an aligned strip of *Graz* including five object classes. Note that the original color coding of the method presented in [Gruber-Geymayer et al., 2005], is adjusted according to our class colors. In addition, we replace the uncertainty regions by an additional class that encodes undefined (white pixels). Table 4.3 summarizes a quantitative evaluation in terms of correctly classified pixels. One can notice that our approach provides improved classification results for *grass*, *tree* and *street*. Due to the large amount of *building* pixels the method in [Gruber-Geymayer et al., 2005] obtains a slightly

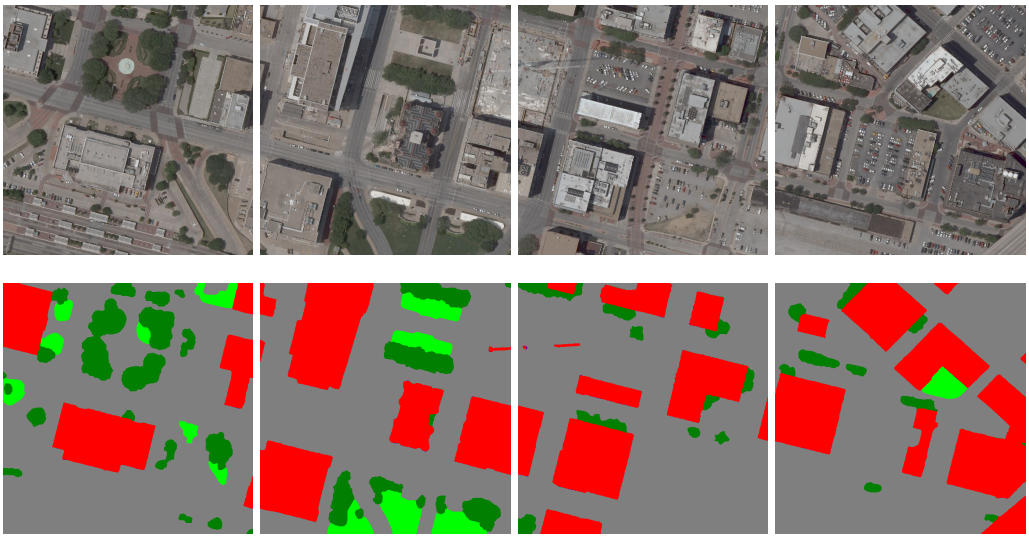


Figure 4.26: A semantic classification obtained for a strip of *Dallas*. The Potts model largely preserves objects boundaries, like building edges, by using a color-driven penalty function. To handle the enormous amount of redundant data we collect the data in 1600×1600 pixels image patches.

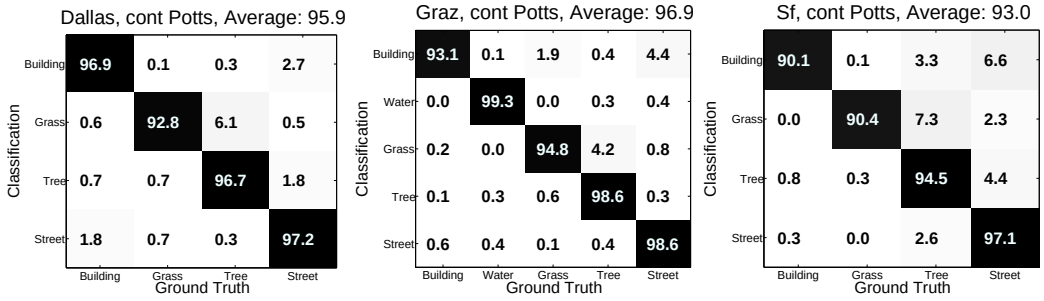


Figure 4.27: Computed confusion matrices for the aerial projects. We obtain classification rates of approximately 90% for the three challenging datasets. The low gray-valued buildings in *Dallas* are sometimes mixed with the *street* class which could be caused by an inaccurate terrain models. Due to similar spectral ranges small shadow regions covering the class *street* are classified as *water* (*Graz*). Many small trees inside of courtyards and the hilly terrain in San Fransisco explain the relatively low classification rates obtained for the *tree* class.

increased averaged accuracy. Aware that the evaluation is far from a fair comparison, we can observe that both methods perform nearly equally although they entirely differ in the used methods and used training data. However, although the methods presented in [Gruber-Geymayer et al., 2005] performs slightly better, our approach provides several

Dallas	*building* [%]	*water* [%]	*grass* [%]	*tree* [%]	*street* [%]	Pixel Avg [%]
Raw	96.20	n.a.	90.57	95.58	96.19	95.61
Potts model	96.91	n.a.	92.81	96.73	97.21	**96.64**
CRF 4-connected	96.99	n.a.	92.60	96.23	97.08	96.54
CRF super-Pixel	97.19	n.a.	92.86	96.53	96.92	96.61

San Francisco	*building* [%]	*water* [%]	*grass* [%]	*tree* [%]	*street* [%]	Pixel Avg [%]
Raw	86.67	n.a.	87.91	91.68	95.39	89.62
Potts model	90.10	n.a.	90.45	94.50	97.11	**92.47**
CRF four-connected	90.06	n.a.	89.69	94.25	97.10	92.34
CRF super-Pixel	89.61	n.a.	91.81	94.90	96.89	92.32

Graz	*building* [%]	*water* [%]	*grass* [%]	*tree* [%]	*street* [%]	Px Avg [%]
Raw	92.98	98.95	93.87	98.00	97.35	95.37
Potts model	93.14	99.29	94.77	98.64	98.59	**95.95**
CRF four-connected	93.15	99.27	94.62	98.48	98.52	95.90
CRF super-Pixel	93.15	99.50	94.79	98.50	98.61	*95.95*
[Gruber-Geymayer et al., 2005]	96.54	99.38	88.95	95.66	97.37	96.34

Table 4.3: A comparison of different refinement strategies: We skip the *water* class for *Dallas* and *San Francisco* due to a missing water areas in the test images. The refinement step successfully improves the final interpretation in terms of correctly classified pixel. In addition, we compare the rates to the ones obtained by a workflow proposed in [Gruber-Geymayer et al., 2005].

advantages. First, our approach provides an interpretation into predefined object classes that are directly learned from the training data annotation. This is useful if *e.g.*training data is automatically derived from web-based GIS systems. In addition, the set of object classes is not limited to a certain group of objects. Second, we directly combine appearance and 3D information, and exploit redundant classifications of perspective input images. This information could be additionally used for different application ranging from automated GIS data derivation to full 3D facade description. Third, our workflow does not introduce static information like that the *building* or *tree* class can be separated from *street* or *grass* by means of the available elevation. It is evident that such an information could be used to improve the final results of our workflow.

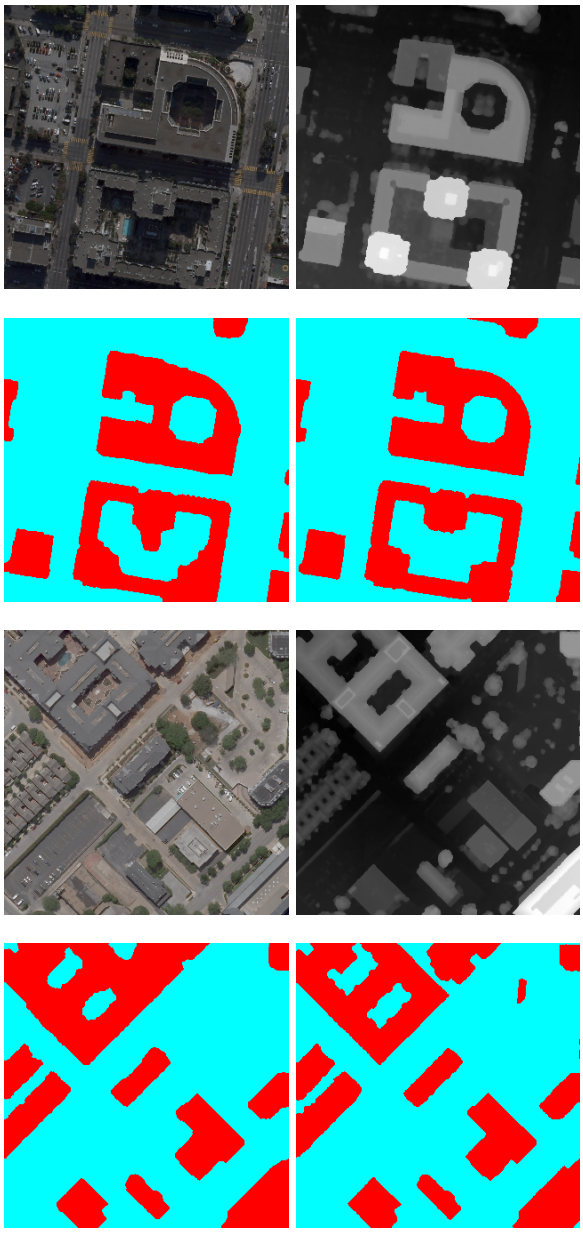


Figure 4.28: Building classification for scenes of *Dallas* and *San Francisco*. We use the edge information, either extracted from the color or the 3D height data, to refine the classification. One can see that 3D height information significantly improves the final building delineation.

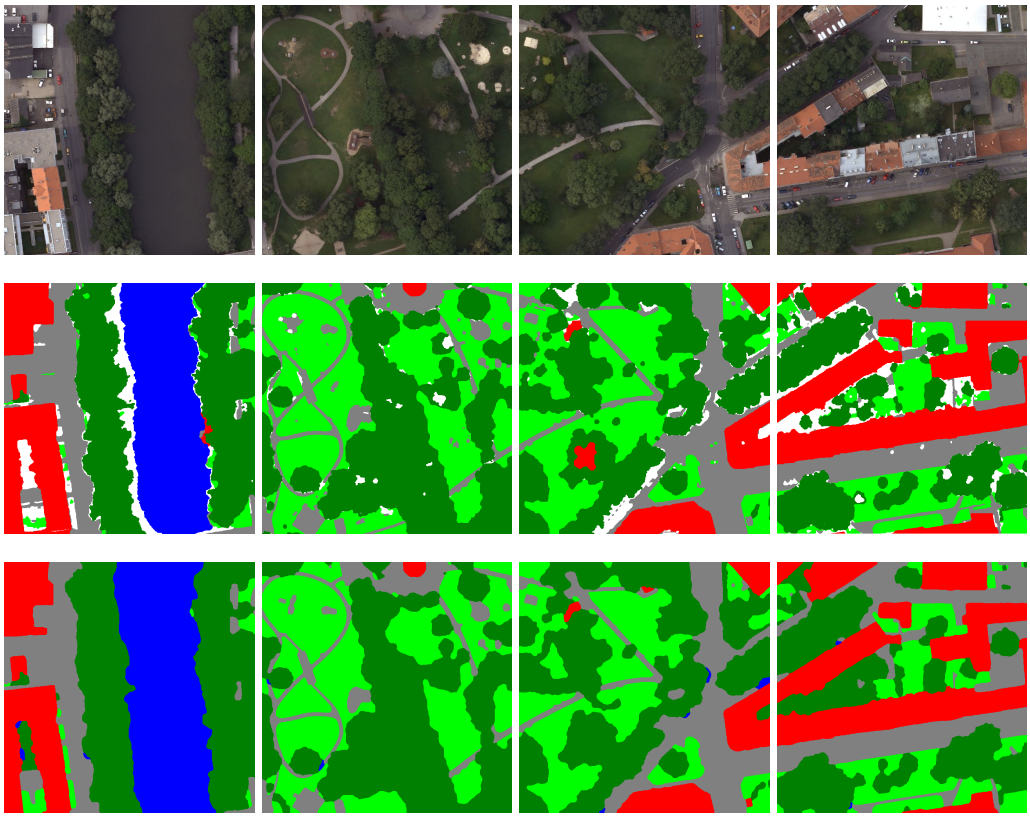


Figure 4.29: A visual comparison between the results obtained with our workflow and the one described in [Gruber-Geymayer et al., 2005]. The first row shows the color image tiles. The results computed with [Gruber-Geymayer et al., 2005] are given in the second row, while our interpretation is depicted in the last row. Note that the refined classification results have been aligned to a common world coordinate system in a first step. In addition, we only compare corresponding object classes, the remaining, mainly uncertainty classes, are skipped for the evaluation procedure and are denoted as white pixels. From a visual inspection one can notice that our approach yields a similar interpretation result without undefined regions.

4.6 Discussion and Summary

This chapter has discussed how to obtain congruent image tiles, represented within a common view. for color, surface models and semantic interpretation. This can be clearly seen as a holistic collection of various scene descriptions. As commonly used in photogrammetry, we utilize the available range information to transform highly overlapping scene observations to an orthographic view. Due to the provided high overlap in the aerial project, redundant measurements for different source modalities are available for each

sampled point on ground. In this chapter the modalities are handled individually to obtain improved fusion results.

We have presented variational approaches for color and height data integration, which can be accelerated using parallel computation. These methods are essential for large-scale computations due to the enormous amount of data. In the experimental section, we have showed that optimization strategies including different types of regularization are important for maintaining the high quality of aerial images within the various fusion steps. In a first step we used the TV-L^1 energy functional, defined over multiple input observations, to form an 2.5D height map of the observed scene. Moreover, we have presented a novel variational method to fuse redundant gray and color images by using wavelet-based priors for regularization. To compute the solution of our large-scale optimization problems we exploit an optimal primal-dual algorithm. We have shown that our fusion method is well-suited for synthetic view generation in high-resolution aerial imagery, but also for an integrated exemplar-based inpainting to remove, *e.g.*, non-stationary objects, like cars.

In order to complete the holistic scene description, we have also aggregated the available multi-view semantic interpretation within the corresponding orthographic view. In the evaluation framework we have investigated how highly overlapping semantic classification into different object classes influences the final result and how optimization techniques can be applied to improve the final class labeling in terms of correctly classified pixels. In particular for the refinement of the accumulated interpretation results, edge penalty functions derived from fused color or height information are utilized within the energy minimization to assign the correct class labels with respect to real object boundaries. Compared to the rates obtained for single images (see Chapter 3), the fusion of redundant initial classifications and a sophisticated refinement stage significantly increase the final percentage of correctly explained pixels to more than 92%. Especially for the building classification we have shown that an integration of 3D information into the refinement step successfully separates the elevated structures from the background.

From this chapter we conclude that redundant scene observations of any modality help to improve the final result concerning noise suppression, the amount of undefined pixels and classification accuracy. Especially, variational methods providing sophisticated regularization techniques, are essential for the processing of aerial imagery due to accuracy, well-defined formulations and efficiency. Moreover, the potential parallelization schemes fit the multi-core technology and enable extremely low computation times, which is important in case of handling immense amount of collected input data.

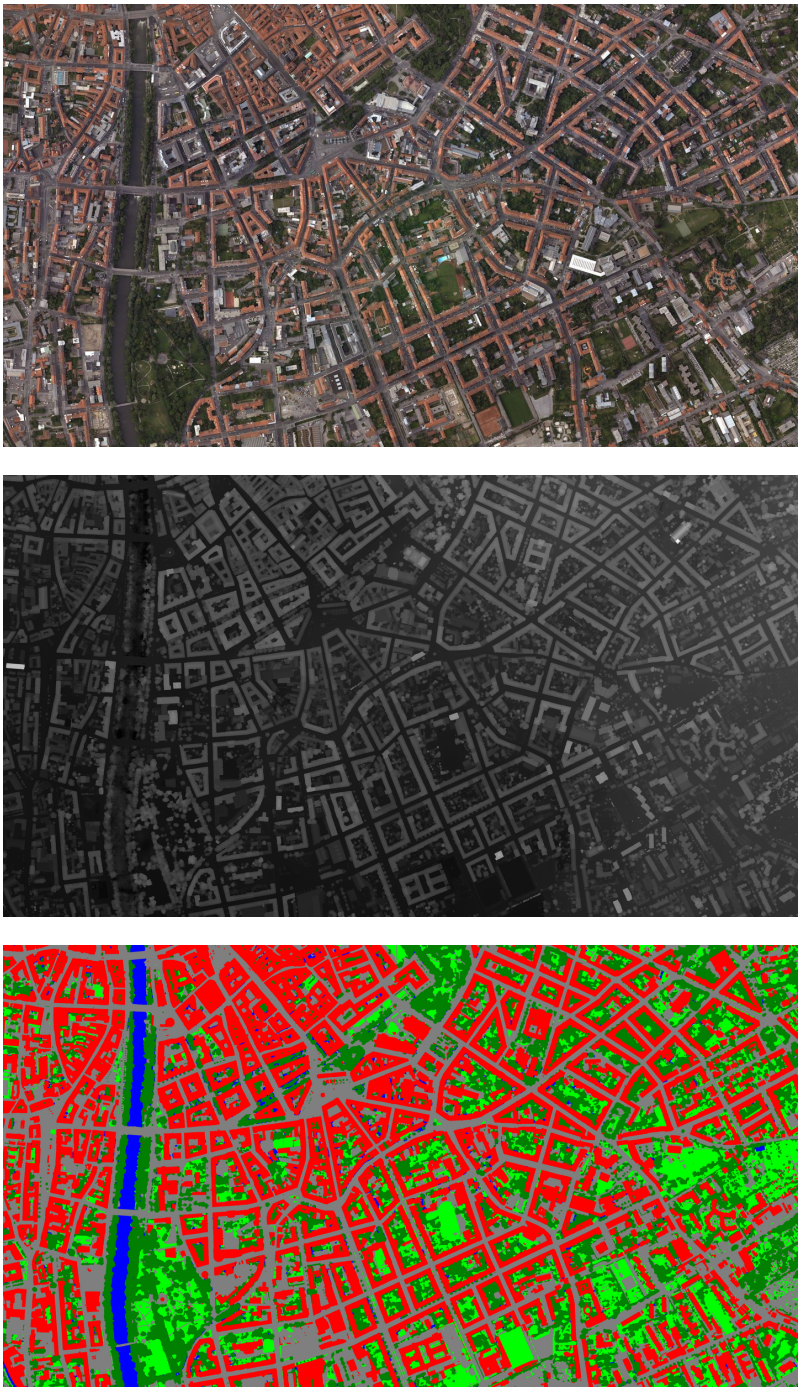


Figure 4.30: A holistic description of *Graz* represented within an orthographic view with a GSD of 8 cm. The proposed workflow provides the fused color data, the surface and terrain model (we only depict the DSM), and a semantic interpretation of the mapped environment by utilizing 155 highly overlapping aerial images. A number of 20×10 patches are combined into individual image structures, each with more than 450 MPixels. The mapped regions shows the inner-city of *Graz* and covers an area of approximately 3.3 sqkm.

Figure 4.31: Inpainting results using a car detection mask. We compute the car mask with the approach described in [Mauthner et al., 2010]. From left to right: The car detection mask, the fusion result computed without using a car detection mask, the result obtained by pure inpainting and the inpainting with supporting exemplars. The car areas are successfully removed in both cases, however the exemplar-based inpainting appears more naturally.

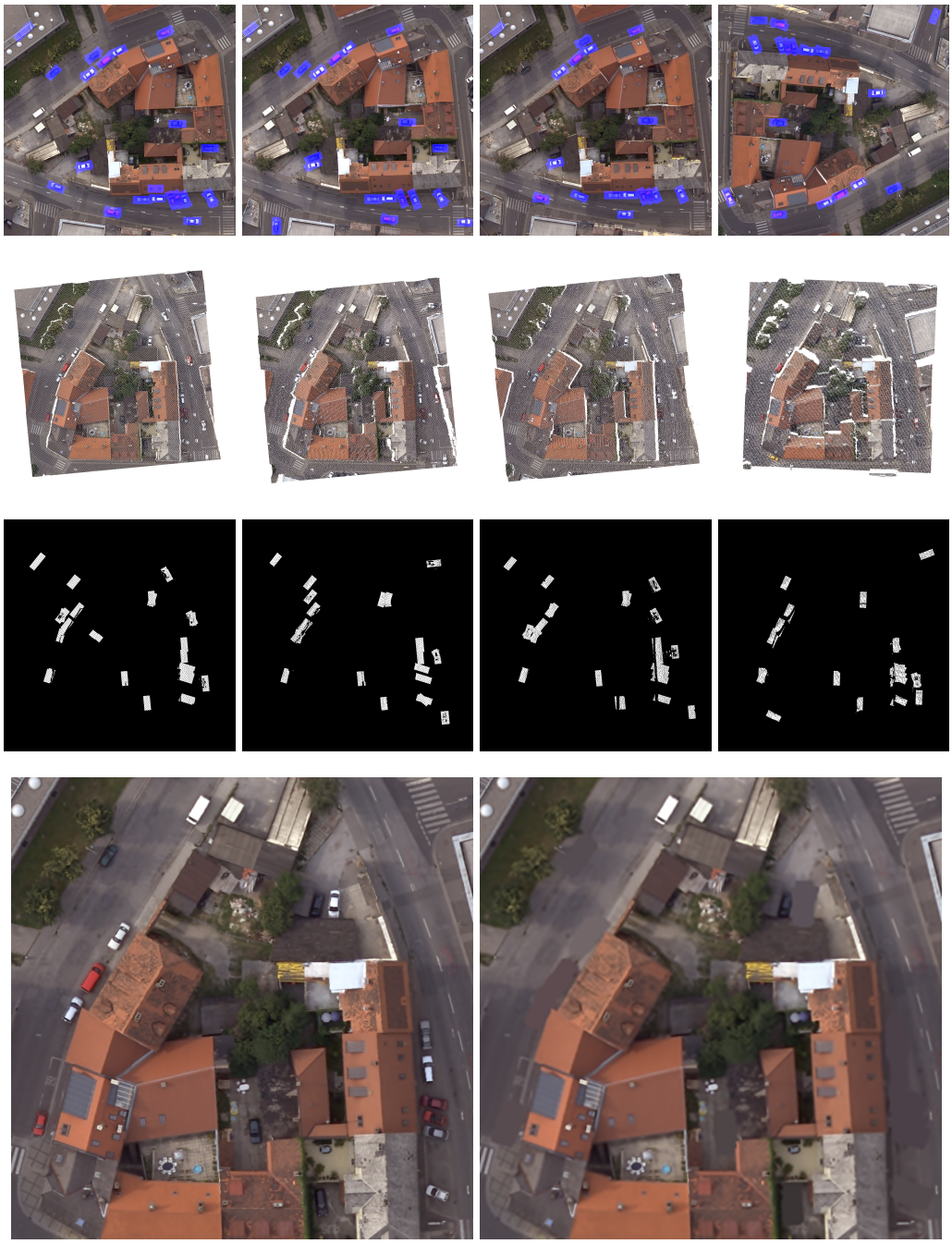


Figure 4.32: Car detection and inpainting from overlapping images. The first row shows some overlapping perspective input images overlaid with the car detection mask. By using the available range images and camera data we project the information to a common orthographic view (the second row shows the color information, while the third row depicts the projected car detection. The fused color image without using the car detection and the inpainting result is given in the last row.

Chapter 5

From Interpreted Regions to 3D Models

In the previous chapters we have highlighted a workflow for digital aerial images, that provide us with a holistic description of urban environments. The processing steps yield many corresponding image tiles, representing color information, surface and terrain models and semantically interpreted regions. This collection of data can be used to generate large-scale synthetic or realistic models of environments in a next step. While photo-realistic city models mainly consist of uninterpreted meshes with assigned texture information, we rather focus on semantic modeling, where the model is constructed according to an available interpretation. These models enable fast search queries and can be efficiently streamed over the Internet or easily applied to procedural modeling. We particularly focus on an efficient and purely image-driven reconstruction of the roof landscape by exploiting the holistic scene description. However, in a similar way the image tiles can also be used to generate a complete street network or to recreate vegetation with compactly parametrized models of trees and bushes. Our approach again extensively integrates both appearance cues and height data to accurately extract building regions and to model complex rooftops. The main processing steps rely on fast color segmentation techniques based on super-pixels, the estimation of suitable geometric prototypes and a quick modeling procedure, where each super-pixel is extruded to 3D by using the available height information. Considering the super-pixels as smallest units in the image space, these regions offer important spatial support for an information fusion step and enable a generic geometrically modeling and a simple description of arbitrary building footprints and rooftop shapes. Again we apply our approach to three different datasets showing environments of *Dallas*, *Graz* and *San Francisco*.

5.1 Introduction

Methods for building detection, extraction and 3D modeling from aerial imagery have become very popular, especially in remote sensing, due to a rapidly increasing number of applications, like urban planning and monitoring, change detection, navigation support, cartography or photogrammetric survey. Large areas of the world are now being mapped at human-scale detail to support these applications. A central task is the detection and 3D modeling of building structures. A comprehensive survey of building detection from aerial imagery is given in [Mayer, 1999]. In recent years an intense research on automatic methods for building modeling at a city-scale has been undertaken [Vosselman and Dijkman, 2001, Parish and Müller, 2001, Werner and Zisserman, 2002, Zebedin et al., 2008, Lafarge et al., 2010]. Nowadays, useful approaches need to be fully automated allowing large-area mapping at low cost. Nevertheless, the enormous amount of collected data requires fast methods and sophisticated processing pipelines, getting by with a minimum of human interaction.

In the case of large-scale datasets the process of building extraction and modeling from urban environments becomes very difficult for many reasons. Buildings are complex objects with many architectural details and shape variations. Buildings are located in urban scenes that contain various objects from man-made to natural ones. Many of those are in close proximity or overlapping and occluding each other, such as parking lots, vehicles, street lamps, trees, etc. Some objects are covered with shadows or cluttered. For instance, Figure 5.1 depicts typical urban scenes taken from three different datasets showing some of these properties. These variations make the task of a general building extraction and modeling challenging.

We therefore propose an approach, which greatly utilizes the available holistic description of scenes, ranging from color and texture, to 3D information and a full semantic interpretation of the pixels. As shown in the Chapters 3 and 4, a sophisticated combination of appearance cues and 3D data provides a reliable detection of building structures in aerial images. Having an accurate knowledge about the land-use enables a model construction according to the corresponding interpretation by simultaneously using available color and height observations.

Additionally, we again introduce an unsupervised color segmentation technique based on super-pixels [Vedaldi and Soatto, 2008] for a generic construction of entire rooftop landscapes. Super-pixels are coherent images regions, describing the smallest unit in the image space and are not limited to predefined sizes or shapes. Hence, in the optimal case a set of super-pixels enables a composition of any building footprint. Together with an estimated plane (here we additionally exploit the congruent height information), each super-pixel is used to form a geometric part of a rooftop. A final refinement step

Figure 5.1: Typical color images of complex urban scenes taken from the dataset *Graz*, *Dallas* and *San Francisco*. Note that each dataset provides an immense amount of building types and variations. In order to handle this challenging data we first propose to utilize elevation measurements in combination with color information and secondly we use color segmentation based on super-pixels to model any type of building shape.

yields a piece-wise planar approximation by taking into account the extracted geometric parametrization of adjacent super-pixels. Therefore, our approach exceedingly exploits the redundancy in the data, mainly given by color, height and building classification, in order to reconstruct the rooftop landscape. Apart from some human interaction to label training maps for learning the classifiers (this has been addressed in the previous chapters), the proposed method runs automatically with a low number of parameters to adjust.

5.2 Related Work

Recent approaches heavily differ in the use of data sources, extracted feature types and the applied models. A couple of recently proposed methods exploit 3D information provided by LiDAR data [Matei et al., 2008, Poullis and You, 2009], but already earlier approaches [Bignone et al., 1996, Cord et al., 1999] used a combination of 2D and 3D information for building extraction and modeling. Matei *et al.* [Matei et al., 2008] proposed a large-scale building segmentation for densely classified urban environments. Their approach accurately preserves object boundaries and enables an accurate modeling of rooftops, however the approach requires a fine parameter tuning. Poullis and You [Poullis and You, 2009] employed a three-staged approach, including preprocessing, segmentation and modeling, for a fully automatic yet large-scale construction of polygonal 3D city models from LIDAR data. They evaluated the approach on a variety of datasets.

Due to improvements in the field of stereo and multi-view matching (see [Hirschmüller, 2006, Irschara, 2011]), an increasing number of methods directly utilize DSMs, that are extracted from redundant images. Lafarge *et al.* [Lafarge et al., 2008] detected rectangular building footprints in the surface models and used symmetry criteria to roughly estimate the geometry of rooftops. In [Lafarge et al., 2010] the authors extended this approach with a library of 3D blocks for improved building generalization from single DSMs. These blocks can be seen as pieces sticked together for building construction and have to be given in advance. A coarse estimation of the rooftop landscape can then be used to detect structures at a finer scale [Dornaika and Bredif, 2008]. In contrast, to exploit a given number of designed models, we rather consider individual image regions, provided by super-pixel segmentation, as the smallest units representing generic building parts.

While Taillandier [Taillandier, 2005] exploited cadastral maps, aerial images and a DSM for a generic modeling, Vosselman and Dijkman [Vosselman and Dijkman, 2001] reconstructed rectangular shaped buildings from point clouds and given ground plans by detecting line intersections and discontinuities between planar faces. Baillard and Zisserman [Baillard and Zisserman, 2000] proposed an automatic method to generically construct a piecewise planar model from multiple images. More generally, Zebedin *et al.* [Zebedin et al., 2008] proposed a concept based on fusion of feature and area information for building modeling. The method relies on directly extracting geometric prototypes, such as planes and surfaces of revolution, taking into account height data, 3D lines and an individual building mask. A CRF-based optimization procedure refines the final result to form consistent and piecewise planar rooftop reconstructions.

Our modeling approach is mainly based on the availability of a semantic interpretation, where building regions can be easily extracted. Related methods involve classifi-

cation techniques to automatically distinguish between mapped objects. Matikainen *et al.* [Matikainen et al., 2007] employed a DSM segmentation and a color-driven classification to discriminate buildings from trees. In [Zebedin et al., 2006] the authors fused information from redundant multi-spectral aerial images within a multi-stage approach to generate orthographic images for color, 3D height and land-use classification.

5.3 Overview

This chapter outlines an approach for efficient, fully image-driven building extraction and synthetic 3D modeling in large-scale aerial imagery, by using an available holistic scene description consisting of color, surface models and semantic interpretation. Our entire modeling concept is summarized in Figure 5.2.

We consider highly overlapping color aerial images and two types of derived height fields as input sources: As described in Chapter 2, a dense matching approach [Klaus et al., 2006] provides us with corresponding range information for each pixel in the input images. Note that the range information defines a DSM in a 3D coordinate system. A DTM, representing the bald earth, is computed in advance from the DSM by using a similar approach as described in [Unger et al., 2010].

In order to compute the building classification we utilize the results of the approach as described in Chapter 3. We perform a binary classification by using the Sigma Points feature representation within RF classifiers in order to obtain an initial interpretation at the pixel level for each image in the dataset. Note that our modeling approach can be fed with any computed semantic interpretation, either the five-class interpretation (*building*, *water*, *tree*, *grass* and *street*), where we finally extract *building* regions, or the pure binary building classification. Due to the high overlap in the aerial imagery, each mapped point on the ground provides class distributions computed from different viewpoints. Then, a pixel-wise fusion step of multiple views into a common 3D coordinate system generates redundant image tiles for various source modalities, like image interpretation, color and height information (see Chapter 4). This fusion step provides corresponding image patches for color, height and semantic (building) classification in an orthographic view. In addition a terrain model enables a direct computation of elevation measurements required for the modeling procedure. Figure 5.3 shows two neighboring sets of the required input data (color, height and *building* class confidences), represented in an orthographic view.

The main step of our approach involves the generic rooftop construction taking into account the computed super-pixels. A generation of super-pixels provides footprints for any object in an observed color image. Taking into account the refined building classification and additional height information, 3D geometric primitives, describing the smallest unit of a building rooftop, can be extracted efficiently. Then, estimated rooftop hypotheses

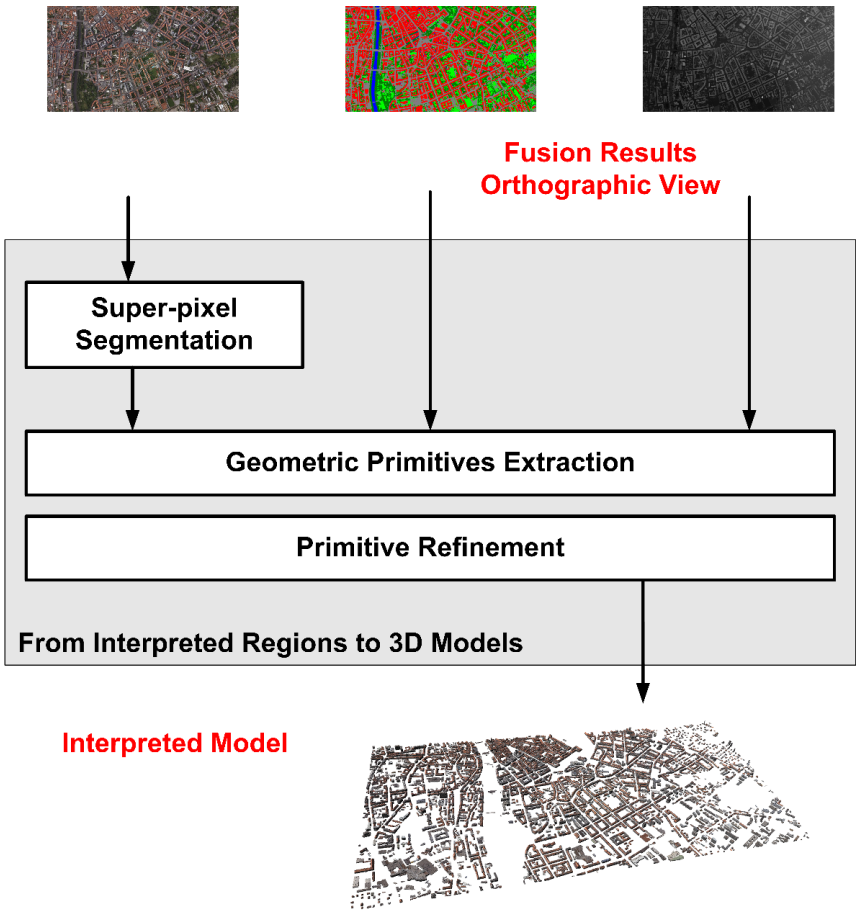


Figure 5.2: Overview of the proposed building extraction and modeling approach: We again utilize color images and height information to detect and construct 3D building models. Note that the building classification and the fusion processing steps are described in detail in the Chapters 3 and 4, respectively.

for each super-pixel in a building (we simply extract connected components on the adjacency graph) are collected and clustered in order to find representative rooftop prototypes. In our case we use a spectral clustering step [Frey and Dueck, 2007] to detect representative geometric prototypes. Finally, a CRF optimization consistently assigns the found prototypes to each super-pixel in building considering the resulting reconstruction error and the neighborhood segments. To show the performance of our modeling approach we apply the method to the datasets *Dallas*, *Graz* and *San Francisco*.

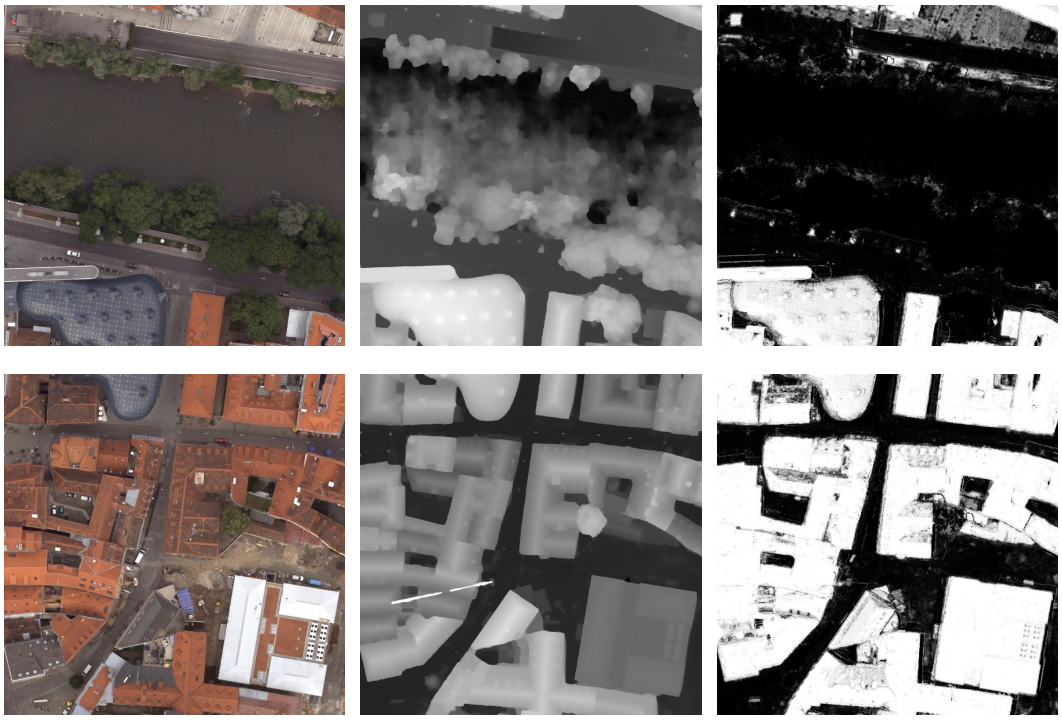
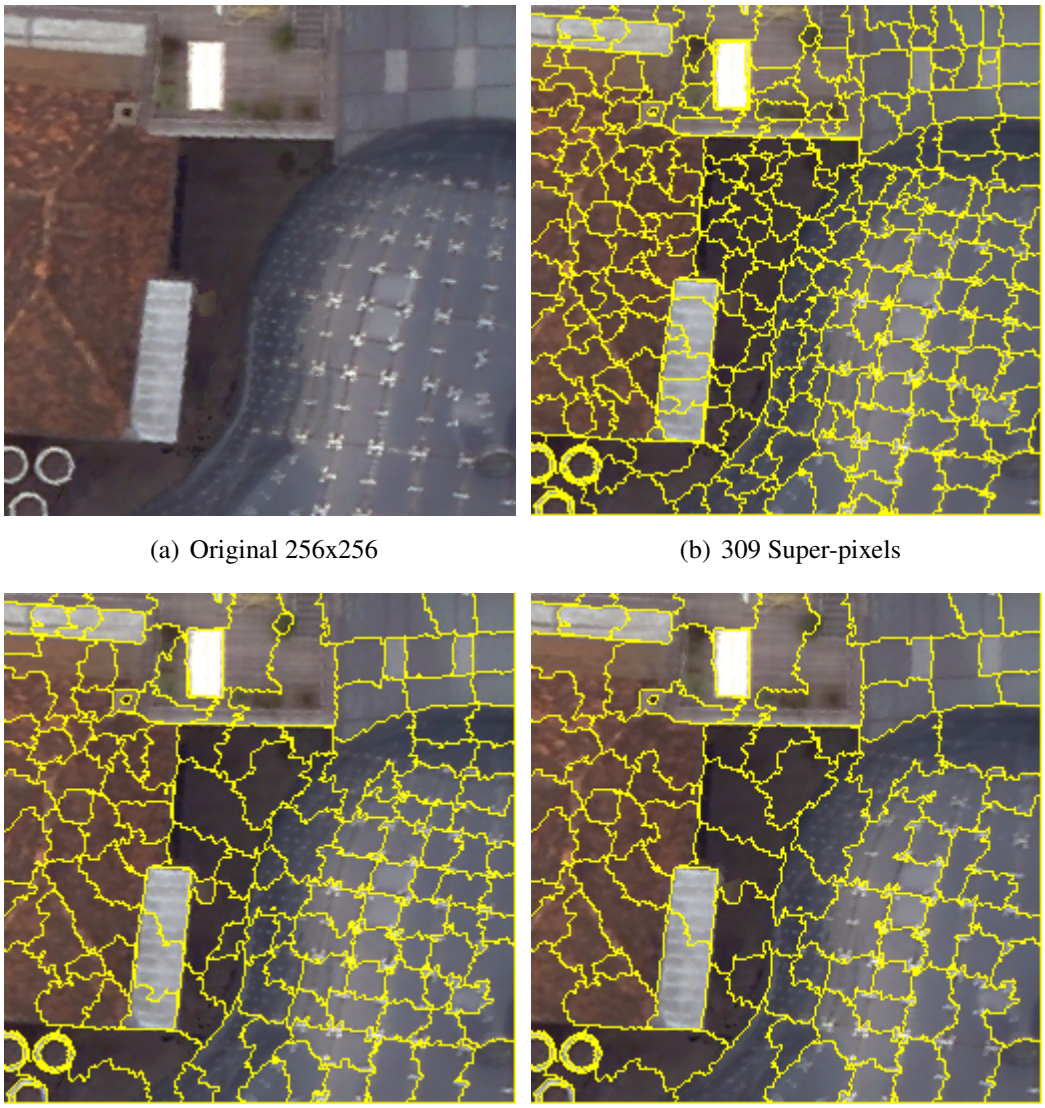


Figure 5.3: Two neighboring sets of input patches of *Graz* used in the modeling pipeline: we use a holistic collection of fused orthographic image tiles consisting of color (first column), height (second column) and raw *building* confidences (last column). Additionally, we exploit the available terrain model to generate the final building models.

5.4 Building Modeling based on Super-Pixels

A variety of modern methods integrate unsupervised image segmentation methods into the processing workflow in order to accurately detect real object boundaries. Several approaches utilize multiple segmentations [Malisiewicz and Efros, 2007, Pantofaru et al., 2008]. However, the generation of many image partitions induces enormous computational complexity and is thus impractical for aerial image segmentation. Fulkerson *et al.* [Fulkerson et al., 2009] proposed recently to use super-pixels, rapidly generated by Quickshift [Vedaldi and Soatto, 2008]. These super-pixels accurately preserve object boundaries of natural and man-made objects. Figure 5.4 depicts a super-pixel segmentation result obtained for different parameter settings.

Applying a super-pixel segmentation to our approach offers several benefits: First, computed super-pixels can be seen as the smallest units in the image space. All subsequent processing steps can be performed on a reduced adjacency graph instead of incorporating the full pixel image grid, thus reducing the data volume. Furthermore, we consider



(a) Original 256x256 (b) 309 Super-pixels

(c) 151 Super-pixels (d) 88 Super-pixels

Figure 5.4: A super-pixel segmentation obtained with Quickshift [Vedaldi and Soatto, 2008] for a small scene of *Graz*. The image partitions are computed for different parameter settings ($\tau = 2.0$, $\sigma = \{8.0, 10.0, 12.0\}$) and result in a different quantity of segments. Note that for all parameters the real boundaries are almost captured completely. In our experiments we keep the parameters fixed for all three datasets.

super-pixels as homogeneous regions, providing important spatial support: Due to edge preserving capability, each super-pixel describes a part of only one class, namely *building* or *background*. Aggregating data, such as the building classification or height information, over the pixels defining a super-pixel compensates for outliers and erroneous pixels.

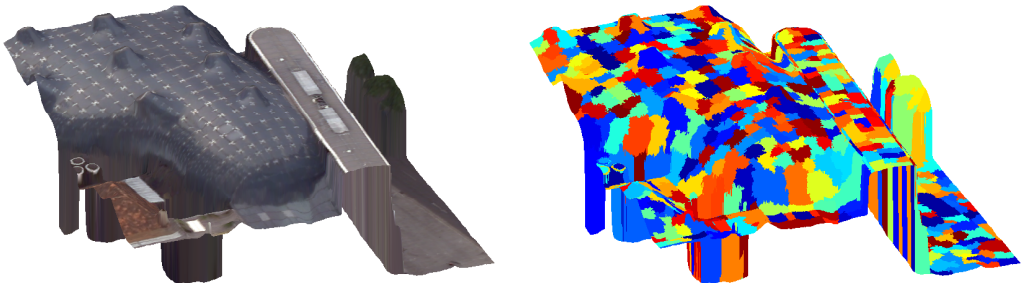


Figure 5.5: A rendering of a small scene of *Graz*. The left image shows the original point cloud rendering using available color and height information. The right figure depicts the rendering overlaid with the Quickshift super-pixel segmentation ($\tau = 2.0$, $\sigma = 10.0$). In our approach we exploit the super-pixel regions to determine representative geometric primitives.

For instance, an accumulation of *building* probabilities results in an improved building classification for each segment. A color averaging within small regions synthesizes the final modeling results and significantly reduces the amount of data. More importantly, we exploit super-pixels, which define parts of the building footprints, for the 3D modeling procedure. Taking into account a derived polygon approximating of the boundary pixels and corresponding height information, classified building footprints can be extruded to form any type of geometric 3D primitives. Thus, introducing super-pixels for footprint description allows us to model any kind of ground plan and in the following the rooftop landscape. In addition, extracted polygons together with determined color and geometric prototypes can be efficiently stored and streamed over the Internet. In Figure 5.5 a point cloud rendering is overlaid with the obtained super-pixel segmentation. Our approach yields a representative geometric primitive for each building super-pixel.

5.4.1 Prototype Extraction

Let S be a set of super-pixels $S_i = \{s_1, \ldots, s_K\}$, classified as parts of a building i, we initially fit geometric primitives to the available corresponding point clouds provided by the fused digital surface model. Due to the fact that a super-pixel is a list of 2D coordinates and that the surface model provides corresponding height information at the pixel level, a terrain model gets directly used to extract a point cloud $P_k = \{\mathbf{p}_1, \ldots, \mathbf{p}_N\}$ in the world coordinate system for each super-pixel $s_k \in S_i$ with $k = \{1 \ldots K\}$ and $|s_k| = N$ individual points. In this work we only consider planes as geometric primitives, however, the prototype extraction process can be extended to any kind of primitives. Due to robustness we apply Random Sample Consensus (RANSAC) [Fischler and Bolles, 1981]

over a fixed number of iterations to find those plane, minimizing the distance to the point cloud, for each super-pixel representing a part of a building. Note that even least square estimates or a fitting with a huge set of pre-defined prototypes could be used instead. This procedure yields an initial rooftop hypothesis $h_k = (\mathbf{q}_k, \mathbf{n}_k)$ for each super-pixel defined by a 3D single point $\mathbf{q}_k \in \mathbb{R}^3$ on the estimated plane and a normal vector $\mathbf{n}_k \in \mathbb{R}^3$. The set of hypothesis H_i for a building i is denoted as $H_i = \{h_1, \ldots, h_K\}$. Figure 5.7 shows a color-coded assignment of individually extracted geometric prototypes.

5.4.2 Prototype Clustering

As a next step, we introduce a clustering of the geometric hypotheses H_i for two reasons: Since the optimization step can be seen as a prototype labeling problem, similar 3D primitives should provide same labels in order to result a smooth reconstruction of a rooftop. Second, the clustering significantly reduces the number of probable labels, which benefits the efficiency of the subsequent optimization procedure. In addition, the space of the possible parametrization can be reduced in advance.

In our approach we use affinity propagation [Frey and Dueck, 2007] to efficiently determine a reduced set $H_i' \subseteq H_i$ of representative exemplars of 3D primitives. Affinity propagation takes as input a distance matrix D of pairwise similarity measurements. Note that the number of exemplars has not to be defined beforehand, but is estimated from the input exemplars. We construct the similarity matrix as follows: For each 3D primitive $h_k \in H_i$, which consists of a normal vector $\mathbf{n}_k$ and a single 3D point $\mathbf{q}_k$ in space, we estimate the reconstruction error by taking into account the current prototype hypothesis h_k and a random set of points P_j extracted from an adjacent super-pixel of s_k. Let be $l = \mathbf{p}_n + t[0,0,1]^T$ the equation of a perpendicular line with $t \in \mathbb{R}$ and let be $\mathbf{q}_k \cdot \mathbf{n}_k + d_n = 0$ the equation of a plane, we can compute a distance d_n between the point $\mathbf{p}_n$ and the point $\mathbf{p}_n'$ projected onto the plane according to

$$d_n = \|\mathbf{p}_n - \mathbf{p}_n'\|_2 = \| \left(\mathbf{p}_n + t[0,0,1]^T\right) - \mathbf{p}_n\|_2 = \|t[0,0,1]^T\|_2 \,, \tag{5.1}$$

with

$$t = \frac{\mathbf{n}_k \cdot (\mathbf{q}_k - \mathbf{p}_n)}{\mathbf{n}_k \cdot [0,0,1]^T} \,. \tag{5.2}$$

Figure 5.6 shows an illustrative example. For each hypothesis h_k we compute a normalized reconstruction error over a randomly sampled point cloud to additionally speed up the computation of the distance matrix required for the clustering procedure. An element (j,k) in the distance matrix, which corresponds to the reconstruction error between two super-pixels j and k is then computed with

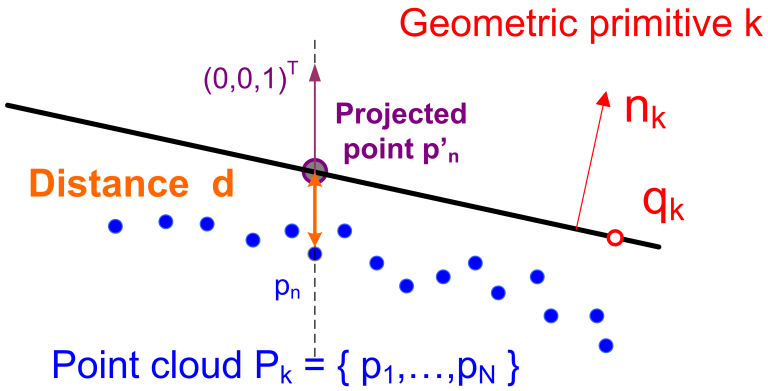


Figure 5.6: An illustration of a distance computation between a geometric primitive defined by a plane $(\mathbf{q}_k, \mathbf{n}_k)$ and a point cloud P_k. We use a random subset of points to estimate a reliable geometric primitive representation of the point cloud provided by each super-pixel.

$$D(j,k) = 0.5 \left(\frac{1}{|n_1|} \sum_{n_1} d^j_{n_1} + \frac{1}{|n_2|} \sum d^k_{n_2} \right) , \tag{5.3}$$

where n_1 and n_2 are indexes of randomly selected points. Since affinity propagation supports sparse similarity matrices [Frey and Dueck, 2007], we consider only adjacent image regions, which additionally reduces computational costs for constructing the distances. The clustering procedure yields a set of representative geometric primitive prototypes, which are used to approximate a rooftop shape with respect to the available height information. Assigned clusters are shown in Figure 5.7 for the *Kunsthaus* scene. One can see that geometrically similar structures (represented within a super-pixel) are assigned a representative geometric prototype. As a next step we use an optimization based on a CRF formulation to obtain a consistent prototype labeling for each individual building.

5.4.3 Prototype Refinement

Although extracting geometric prototypes by using super-pixels captures some local information, the super-pixel regions in the image space, assigned with geometric primitives, are handled nearly independently. In order to incorporate spatial dependencies between nodes defined on the image grid, CRF formulations [Boykov et al., 2001] are widely used to enforce an evident and smooth assignment of probable label candidates. In contrast to minimizing the energy on a full four- or eight-connected image grid we apply the CRF stage defined on the super-pixel neighborhoods similar as proposed in [Fulkerson et al., 2009]. This optimization provides a consistent labeling of the geometric prototypes in

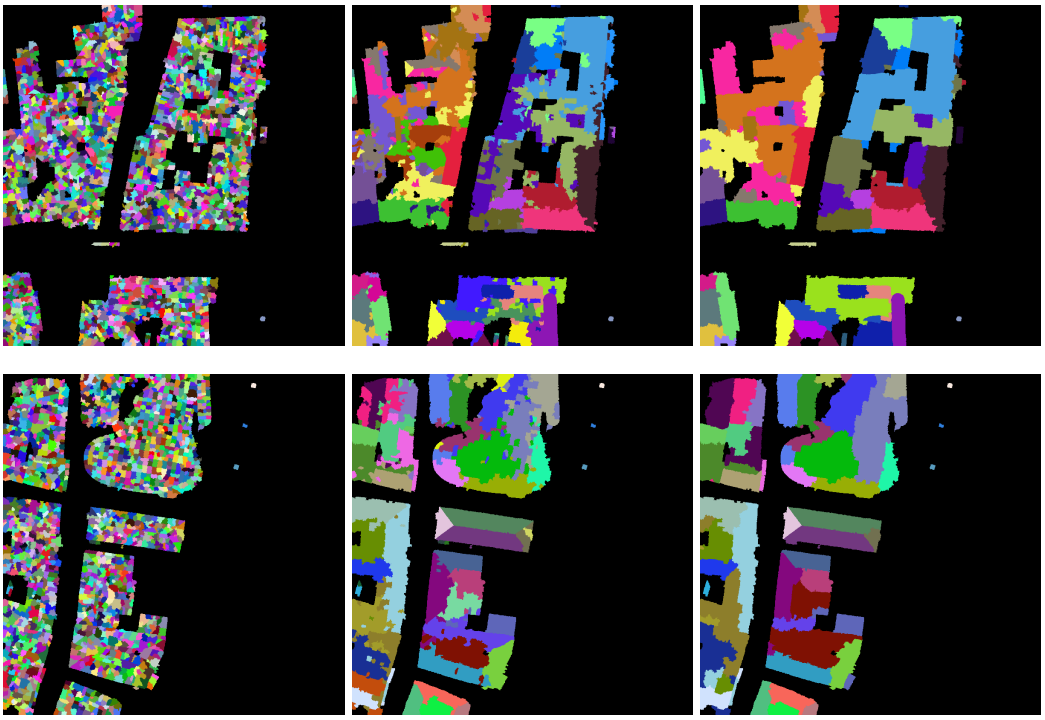


Figure 5.7: Intermediate results for the *Kunsthaus* scene in *Graz*. The first column depicts the computed super-pixels overlaid with the building mask (see Section 5.4 and 5.4.1). Note that each super-pixel is assigned an individual geometric prototype. The second column shows the assignment after clustering the geometric prototypes. The result of the CRF-based refinement step, which groups super-pixels by taking into account the geometric primitives, is given in the last column.

order to obtain piecewise planar rooftops.

In our case **c** represents a labeling of the adjacency graph that assigns each super-pixel s_k in a building i a final label $c_k \in H'_i$, where H'_i is the set of representative geometric prototypes obtained by the clustering process. In this case a label is a possible assignment to a specific geometric primitive.

Let $G(S, E)$ be an adjacency graph with a super-pixel node $s_k \in S$ and a pair $(s_k, s_j) \in E$ be an edge between the segments s_k and s_j, then an energy can be defined with respect to the class labels **c**. Generally, the energy can be defined as

$$E(\mathbf{c}|G) = \sum_{s_k \in S} D(s_k|c_k) + \lambda \sum_{(s_k,s_j)\in E} V(s_k, s_j|c_k, c_j) \;, \tag{5.4}$$

where $D(s_k|c_k)$ expresses the unary potential of a super-pixel node. Similar as proposed in [Zebedin et al., 2008], the unary potential $D(s_k|c_k)$ denotes the costs, in terms of

summed point-to-plane distance measurements, of s_k being assigned a geometric prototype. For each geometric primitive in $h_j \in H_i'$ we compute the unary potential for a super-pixel s_k with

$$D(s_k|c_k) = \frac{1}{|n_1|} \sum_{n_1} d_{n_1}^j \ . \tag{5.5}$$

where n_1 is a random subset of points and $d_{n_1}^j$ is the point-to-plane distance as defined in (5.1). In order to obtain a smooth geometric prototype labeling within homogeneous building areas, we compute the pairwise edge term $V(s_k, s_j|c_k, c_j)$ between the super-pixels s_k and s_j with

$$V(s_k, s_j|c_k, c_j) = \frac{b(s_k, s_j)}{1 + g(s_k, s_j)} \delta(c_k \neq c_j). \tag{5.6}$$

The function $b\,(s_k, s_j)$ computes the number of common boundary pixels of two given segments, $g(s_k, s_j)$ is the L^2 norm of the mean color distance vector and $\delta(\cdot)$ is a simple zero-one indicator function. The scalar λ controls the influence of the regularization and is estimated by using cross validation. In this work we again minimize the energy defined in (5.4) by using α-expansion moves [Boykov et al., 2001]. A refined labeling of prototypes is shown in the right column of Figure 5.7.

5.4.4 Rooftop Modeling

So far, the footprint of each building consists of a set of defined super-pixels in the image space. In order to obtain a geometric footprint modeling of each super-pixel, we first identify common boundary pixels between adjacent building super-pixels. For each super-pixel, this procedure extracts a specific set of boundary fragments, which can be individually approximated by straight line segments. A pairwise matching of collected line segments yields a closed yet simplified 2D polygon. Note that we assume convexity of these polygon in order to obtain a quick triangulation. Taking account of the DTM and the refined geometric primitive assignment, the footprint polygons defined by a number of vertexes are then extruded to form small units of a rooftop: distinctive 3D rooftop points are determined by intersecting the plane (given by the geometric primitive) with a line, directed to $[0, 0, 1]^T$, going through the corresponding vertex on ground. For the purpose of visualization, we simply use a 2D Delaunay triangulation[1] to generate the models of the buildings. An individual 3D building model of our approach can now be seen as a collection of composed building super-pixels having identical building and rooftop prototype indexes, respectively. It is obvious that a hierarchical grouping of super-pixels could

[1] CGAL, Computational Geometry Algorithms Library, www.cgal.org

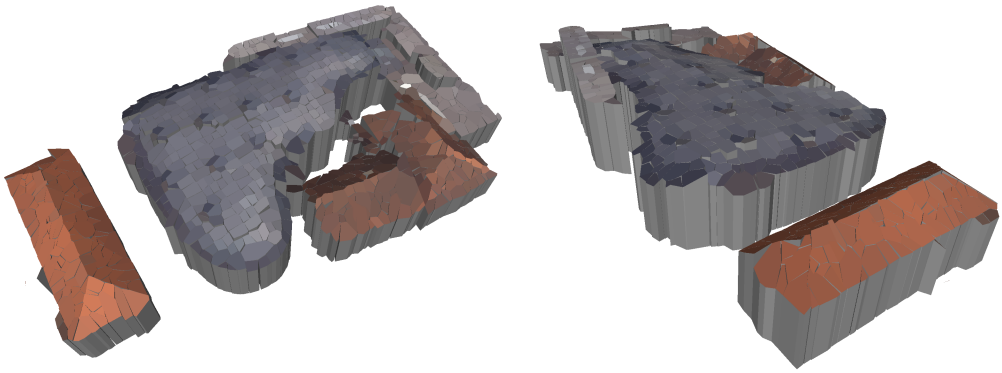


Figure 5.8: The resulting 3D model for the *Kunsthaus* and a neighboring building. Each building consists of a collection of super-pixels, represented as polygons, with an assigned mean color value and a geometric prototype. It can be noticed that the red rooftop in the foreground is reconstructed accurately. The optimization process results in 4 dominant geometric prototypes. In the case of the complex bluish structure, it is obvious that the resulting model is composed of many (planar) geometric primitives.

be used to further simplify the resulting building model. Figure 5.8 displays the result of our modeling procedure, obtained for the *Kunsthaus* scene.

5.5 Experiments

This section evaluates our proposed concept on a large amount of real world data. Thus, we present results of our building generalization and perform quantitative and visual inspection of the constructed models for the three datasets *Dallas*, *Graz* and *San Francisco*. Due to limitations in memory the building model is computed for image tiles with a size of 1600×1600 pixels. Table 5.1 summarizes the basic information for the three datasets used in the evaluation.

In our performed experiments, Quickshift is applied to a five-dimensional vector, consisting of pixel location and CIELab color. The parameters for Quickshift are set to $\tau = 2.0$ and $\sigma = 8.0$. It turned out that these parameters capture nearly all object boundaries in some observed test images. In addition, this parameter setting has given a reliable trade-off between generating sufficiently small regions in order to accurately reconstruct curved shapes at low computational costs. For a 1600×1600 pixel image tile we obtain approximately 4800 super-pixels with nearly constant regions sizes.

In order to obtain a quantitative evaluation, the root mean squared error (RMSE) over

Dataset	Image tiles	Area	# of Buildings	# Pixels
Dallas	11×11	7 sqkm	1160	76.81e6
Graz	24×14	5.5 sqkm	2572	298.6e6
San Francisco	11×11	7 sqkm	1927	113.91e6

Table 5.1: Basic information for the evaluation. We perform the experiments on the full area covered by the provided perspective aerial images. Due to high redundancy in the imagery an image tile includes information from up to 15 perspective views and has a dimension of 1600×1600 pixels, which corresponds to a real size of 240×240 meters for *Dallas* and *San Francisco*, and 128×128 meters for *Graz*, respectively.

all building pixels is computed between the DSM values and the heights obtained by our 3D modeling procedure. Since there exists no ground truth data, *e.g.*, in form of separately generated LiDAR point cloud we have to assume that the original point cloud, provided by the fused digital surface model, provides us with ground truth heights. Clearly, image regions with large reconstruction errors caused by partial occlusions or specular facades will distort the error measurements to some extent. For each building separately we determine the RMSE between the 3D points, resulting from the dense matching and the fusion step, and the corresponding (refined) geometric primitive. To obtain a meaningful error computation we compute median, first and second order statistics over all buildings in a dataset. In addition, we repeat the experiments for different λ, which controls the trade-off between model simplification and data fidelity. Table 5.2 summarizes the obtained results of our quantitative evaluation. We can observe that our modeling approach obtains an acceptable small deviation for *Graz* and *San Francisco* with lower than 1.5 m. Taking into account the obtained results for *Dallas*, one can notice that reconstruction errors caused by specular facades and huge areas of occlusions drastically corrupt the accuracy of our modeling process in terms of an increased RMSE and high standard deviation. Moreover, it is obvious that the intensity of the regularization introduces higher deviations between the real and geometrically refined point cloud. In addition, it is evident that the strength of regularization controls the model simplification, and thus, the number of residual geometric prototypes. While no regularization yields an averaged number of 24 extracted geometric candidates per building block (*Graz*), a $\lambda = 1.0$ reduces to a quantity of 11 prototypes. Figure 5.9 depicts a scene of *Graz* from two viewpoints for different values of regularization. Furthermore, it is obvious that the model computed with $\lambda = 0.01$ shows high granularity with respect to the extracted prototypes, while the result obtained with $\lambda = 5.0$ appears too simplified. Therefore, we choose values, ranging from 0.1 to 1.0, to construct the building models. Note that even an adaptive value can be used by incorporating the original point cloud into an efficient feedback loop.

	Dallas			*Graz*			*San Francisco*		
λ	Median	Mean	Std	Median	Mean	Std	Median	Mean	Std
	[m]	[m]	[m]	[m]	[m]	[m]	[m]	[m]	[m]
0.001	1.159	2.016	3.404	0.963	1.047	0.688	1.212	1.452	1.136
0.01	1.178	2.024	3.423	0.960	1.051	0.690	1.197	1.449	1.137
0.1	1.153	2.028	3.404	0.960	1.056	0.692	1.220	1.455	1.130
1.0	1.398	2.213	3.395	1.097	1.191	0.708	1.423	1.649	1.140
5.0	2.113	2.835	3.508	1.720	1.773	0.939	1.868	2.082	1.320

Table 5.2: Statistics of the error measurements between the (real) DSM values and reconstructed heights using our modeling approach. We calculate the deviations in meters over all buildings in datasets for different strengths of regularizations.

Average processing time per image tile (1600×1600 pixels)	65 s
Segmentation	**25.3** %
Extract super-pixel information	6.2 %
Prototype Extraction	**19.4** %
Prototype Clustering	**47.7** %
Prototype Refinement	0.6 %
Rooftop Modeling	0.7 %

Table 5.3: Time consumption of the individual processing steps for a single image patch. The super-pixel segmentation, the prototype extraction and clustering consume most of required computation time. These values can be significantly reduced by, *e.g.*, implementing the proposed steps on a GPU.

In Figures 5.10, 5.11, 5.12 and 5.13, computed 3D models are shown for *Dallas*, *Graz* and *San Francisco*. For efficiency and large-scale capability, these models are sticked together in tiles of 1600×1600 pixels. Given the fused color, height and classification images, the entire 3D model of *Graz* can be computed within a couple of hours using a standard dual-core machine. Table 5.3 outlines an assembly of the time consumption for each processing step. Super-pixel segmentation, prototype extraction and clustering consume most of the required computation time.

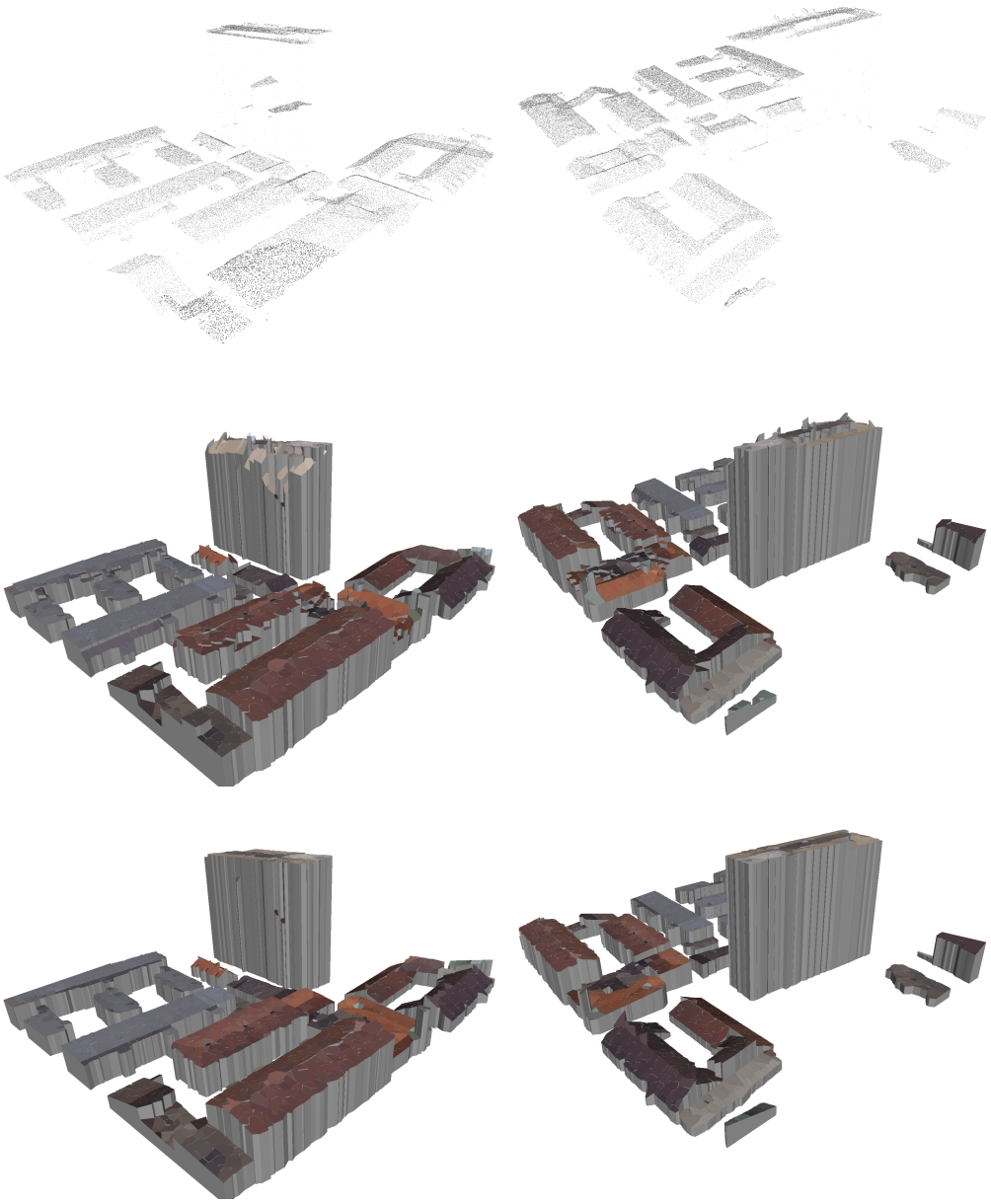


Figure 5.9: 3D building models of the *Griesplatz* (located in *Graz*) shown from two different viewpoints. The model is computed for different intensities of regularization. From top to bottom: point cloud provided by the DSM, model refined with $\lambda = 0.01$ (the model appears very rough) and $\lambda = 5.0$ (the model seems over-smoothed). Thus, we choose the regularization between 0.1 and 1.0.

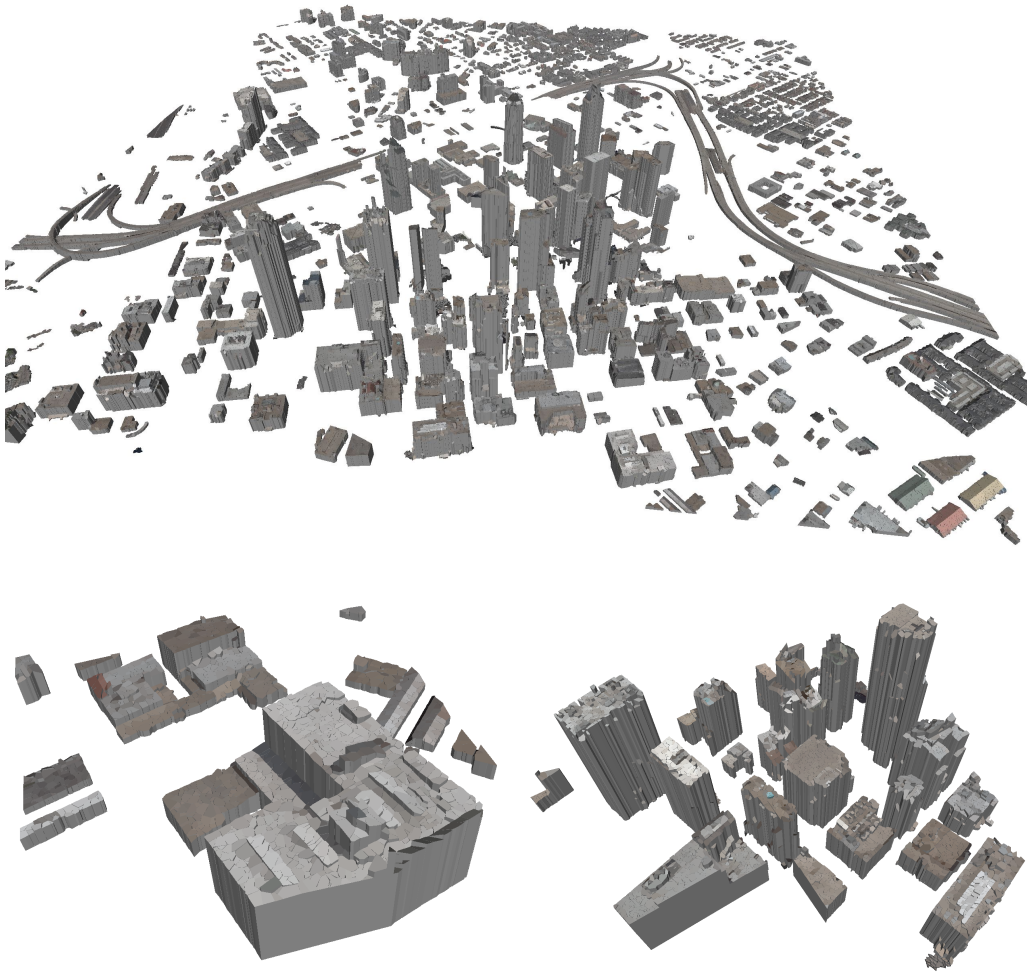


Figure 5.10: 3D building models of *Dallas*. The full model of *Dallas* (first row) covers an area of approximately 7 sqkm and can be constructed within a couple of hours on a standard machine. In particular, computed building models of skyscrapers suffer from reconstruction errors, caused by massive occlusions and large specular facades. To be optimistic, an improved multi-view matching method [Irschara, 2011] will overcome problems, like an inconsistent prototype assignment. Moreover, one can see that elevated circulation spaces, such as highways, are classified as building structures and therefore modeled with the proposed pipeline.

5.6 Discussion and Summary

In this chapter we have described an efficient, purely image-driven approach for constructing semantic 3D models of buildings by utilizing an available holistic scene description consisting of color, 3D height information and a corresponding semantic interpretation.

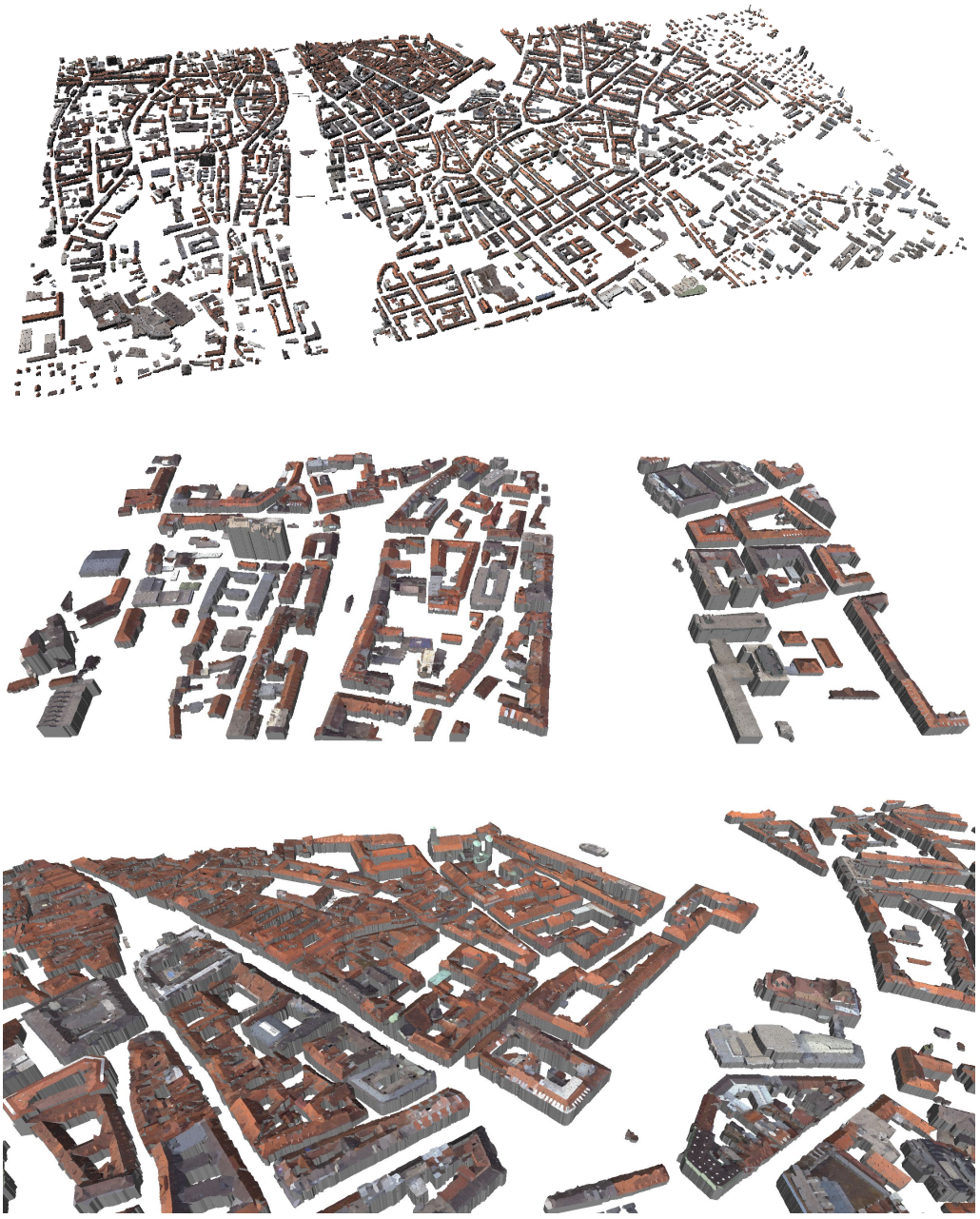


Figure 5.11: The 3D building models of *Graz* covers an area of approximately 5.5 sqkm. The model is depicted from different viewpoints. Note that even complex building structures are preserved accurately.

In particular, involving a super-pixel segmentation enables a generic 2D modeling of any building footprint and in the following a modeling of complex rooftop shapes. Since the super-pixels describe the smallest units in the image space, processing steps, like geo-

metric prototype extraction and refinement, benefit from a reduced image grid and thus significantly reduce the computational complexity. Hence the holistic description within the orthographic view is extended with super-pixel segmentation results in order to enable efficient processing steps, like the street network extraction or an alignment with web-based GIS data, on an adjacency graph. We have applied our approach to different aerial projects and have demonstrated large-scale capability with low time consumption. In contrast to photo-realistic mesh-based models, building models at a city-scale, constructed with the proposed methods, are comprised of a huge collection of uniquely identified building blocks, where each block is composited as an assembly of grouped pixels (in our case super-pixels). Each super-pixel is assigned with derived properties, like a geometric primitive, a mean color, an approximated boundary, etc., which can be efficiently represented in a hierarchical data structure. This hierarchical structure can be easily exploited to efficiently generate models at different levels of detail with CityGML[1] or to provide input sources for grammar-based or procedural visualization methods, such as CityEngine [Parish and Müller, 2001]. It is obvious that an optimization strategy would additionally reduce the complexity of the generated building models by collecting sets of super-pixels that describe basic 2D shapes, like rectangles or circles.

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

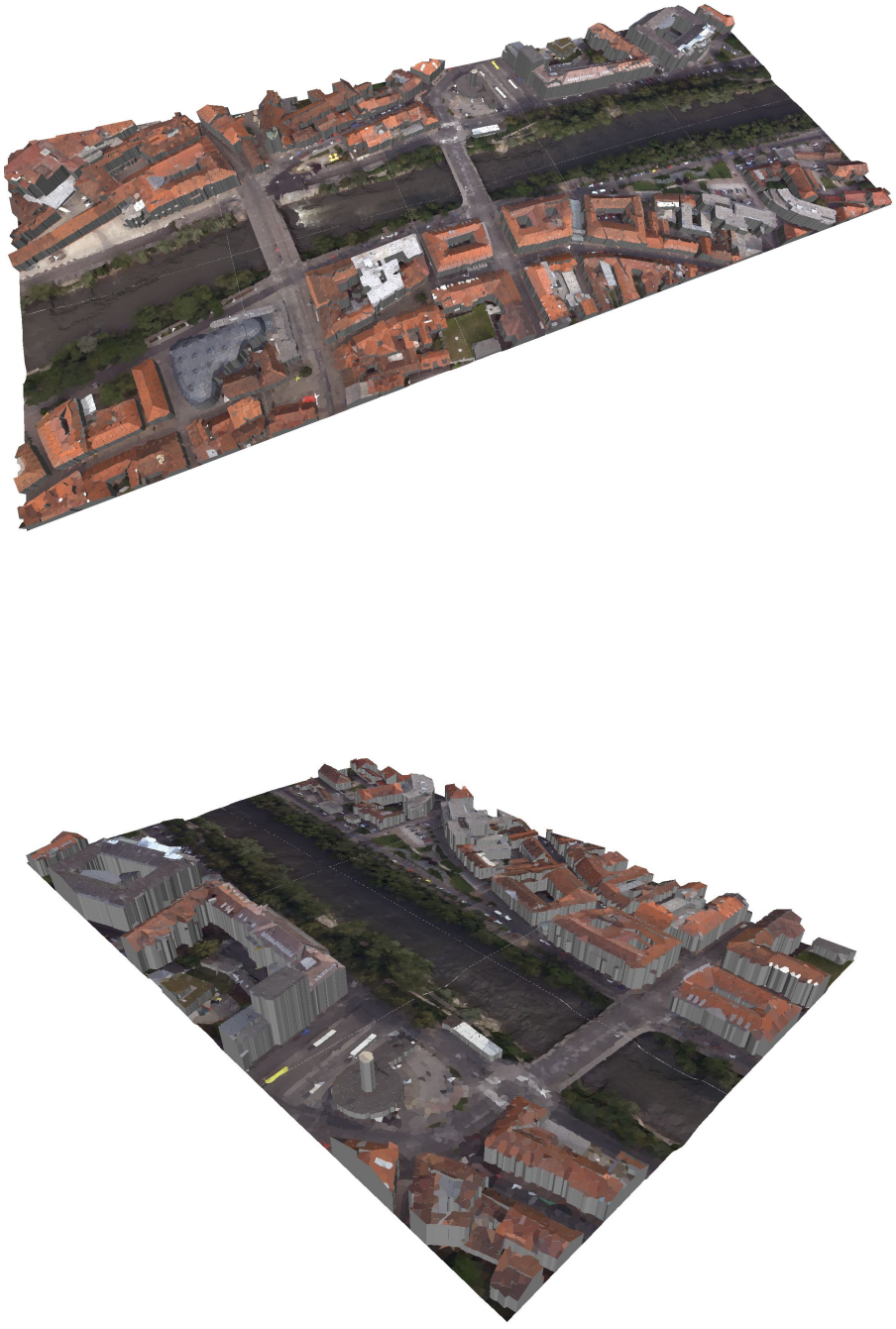


Figure 5.12: 3D building models of *Graz*. We additionally overlaid the 3D visualization with a triangulated DTM. One can observe that this background information improves the impression of the model nicely. Note that every building is identified by an unique key and fully described with a few parameters.

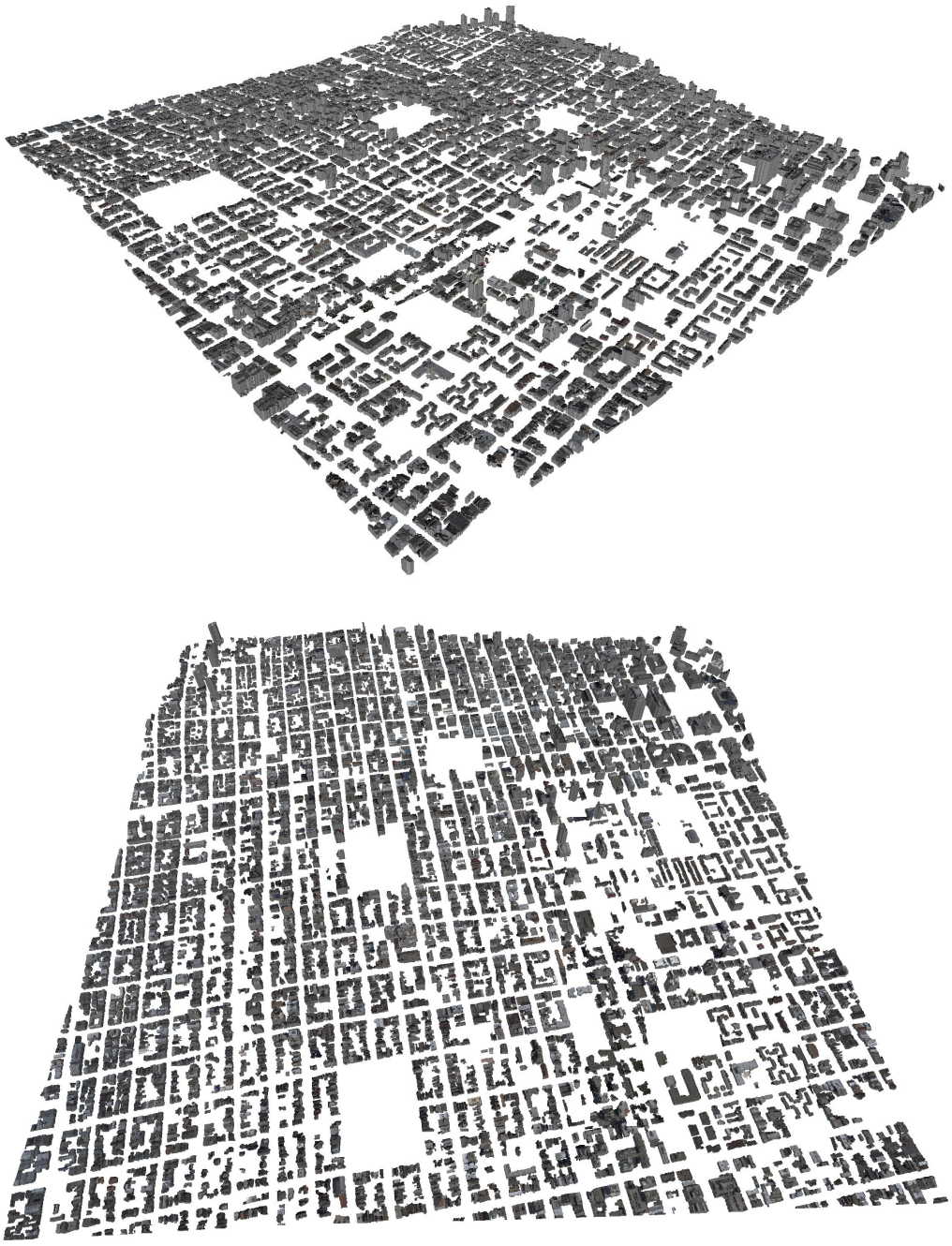


Figure 5.13: 3D building models of a part of *San Francisco*. The model of *San Francisco* covers an area of 2500×2500 meters. Since our holistic collection also offers a fused terrain model, the building block can be placed according to the provided real height values.

Chapter 6

Conclusion

This thesis has addressed the problem of semantic image interpretation of large-scale aerial images, utilizing redundancy, appearance and 3D information. The redundancy in the aerial data, which results from the highly overlapping image acquisition, enabled image-based methods for the estimation of dense and pixel-synchronous 3D scene information.

Based on a novel statistical descriptor and a fast classifier, we have derived a semantic interpretation workflow that compactly combines redundant measurements for each pixel, consisting of color, derived edge responses and height values. The experiments have shown that the compact integration of these measurements improves the classification accuracy significantly. This has also confirmed to use sophisticated image representations, which tightly integrate appearance and 3D information. Applying the workflow to every image in the aerial data has obtained a redundant image interpretation and can thus be seen as a large-scale multi-view classification strategy.

Since the full scene geometry is given by the 3D reconstruction, a fusion step has been introduced to integrate the highly overlapping and therefore highly redundant information into a common view. A projection to this common view has permitted the fusion of multiple observations for color, height and semantic classification by using powerful energy optimization methods. In particular, energy optimization methods, defined over multiple observations, have yielded spatially consistent and noise-reduced results. Undefined regions, mainly caused by moving objects or occlusions, are successfully filled by considering highly redundant image data. From the experiments we can conclude that task-specific regularization strategies, such as total variation, wavelets or the Potts model, are essential to integrate multi-view information for color, object class assignments and height values.

The derived holistic scene description has finally enabled a computation of parametrized building models - a step towards a fully interpreted virtual city. The building models, in

turn, have been computed through utilizing the interchange between appearance and 3D information cues.

By using the three challenging datasets of *Dallas*, *Graz* and *San Francisco*, we have demonstrated a high generalization capability. Additionally, only a few parameters, mainly chosen from reasonable value ranges, are required to control the entire workflow. In the following section we summarize the proposed methods and outline some ideas for future research.

6.1 Summary

The presented workflow in this thesis takes as input redundant color and range images and generates a set of congruent images. They consist of fused color information, surface models and a refined semantic labeling, which assigns a specific object class to each pixel. In this work we have addressed the problem of building classification, as well as the multi-class case, where we distinguish between the classes *building*, *street*, *tree*, *grass* and *water*. In order to obtain a reliable semantic labeling we propose to use available appearance and 3D information cues extensively.

In Chapter 3 we have worked out the advantages of digital aerial mapping and have described the required source modalities, like the generation of redundant range images and the computation of terrain models. Additionally, this chapter has outlined the orthographic representation and the characteristics of the applied real-world aerial projects.

In Chapter 3 we have introduced the semantic interpretation of single high-resolution aerial images. To obtain a reliable object class assignment for each pixel, we have outlined a novel region descriptor which is based on Sigma Points features. This representation allows us to integrate appearance and 3D information compactly, without considering normalization issues. The use of efficient data structures, such as integral images and fast multi-class learners, offers low computation times. In addition, we have outlined the utilization of unsupervised segmentation, providing important spatial support and optimization strategies to accurately delineate the mapped objects with respect to the real boundaries. The per-image interpretation workflow has been evaluated on standard benchmark datasets and huge aerial images. On aerial images we have demonstrated that a tight combination of appearance and 3D information is essential for an accurate semantic interpretation at the level of pixels.

Chapter 4 has studied the fusion of redundant observations for the obtained semantic interpretation, color and height information. Due to the high geometric redundancy of the aerial imagery and the availability of well-defined scene geometry, a projection to a common view has provided highly redundant measurements for each involved modality. We have thus demonstrated that available range images, representing the scene geometry,

can be successfully used to align the overlapping perspective image information into a common view. In a first step energy optimization has been used to integrate redundant scene observations for color and height. We have applied the well-established TV-L^1 model, as well as an extended fusion model that uses improved wavelet-based image priors. To obtain a congruent semantic interpretation in the orthographic view, the highly redundant multi-view classification has been projected to the orthographic view and then been aggregated. An optimization based on the Potts model has offered a simple way to capture object edges by incorporating appearance or height data. The evaluation has clearly shown that the geometric, but also the radiometric redundancy, is important for the generation of a holistic scene description composed of color, height and semantic interpretation.

Finally, in Chapter 5 we have addressed the modeling of the obtained semantic interpretation. In contrast to photo-realistic modeling, where the models usually consist of textured meshes, we have focused on virtual modeling providing semantic knowledge about the underlying content. We have particularly discussed the construction of fully parametrized 3D building models. The proposed approach has again relied on an interaction between semantic classification, available color and height information. Our approach has obtained fully parametrized building models. The estimated parameters have then been used for 3D visualization, consisting of a data-efficient assembly of polygons, geometric prototypes and mean color values.

6.2 Outlook

This thesis has described an approach with a high degree of automation to derive a holistic description of aerial images. Although we have presented the work in the context of aerial imagery, it is evident that the proposed concepts could also be applied to terrestrial images or any other scale of image acquisition. In some points, there is still room for improvements in further research.

6.2.1 Training Data

A big step towards an automatic image understanding must be the reduction of massive human interaction that is required to generate training maps providing the object labels. While web-based photo communities on the Internet, like *Flickr*, *Google Images*, etc., which offer multiple images from nearly every hot-spot in the world, are usually exploited to improve the 3D reconstruction for free, the acquisition of labeled data for a semantic classification can be accomplished by playfully contributing a lot of manual interaction, such as with *LabelMe* [Russell and Torralba, 2009] or by simply using ex-

pensive paid services, like *Amazon Mechanical Turk* [Sorokin and Forsyth, 2008]. One goal must therefore be to take advantage of GIS resources, in form of publicly available user-annotated Internet data, in order to reduce the manual intervention and to improve the semantic classification. The *OpenStreetMap*[1] initiative particularly offers a rapidly growing amount of vector-based data that contains GPS locations of buildings, sights and trees, courses of rivers and streets. Taking into account this information, annotations for object classes, like buildings, trees, streets, etc. might be extracted, forming an enormous set of training labels at the pixel-level for each dataset without considering the coverage and landscape characteristics.

6.2.2 Joint Estimation of 3D and Semantic Interpretation

In this thesis we have assumed that 3D scene structure is computed in advance from overlapping source images. Although we have proposed to compactly combine available appearance and 3D information for an accurate semantic interpretation, one might think that classification and 3D reconstruction might be estimated jointly, since these problems are mutually informative. On the one hand, height data can be a very informative cue to discriminate between object classes, like building and streets, or to distinguish between different types of vegetation. On the other hand, information about the objects yields a strong prior about the underlying 3D structure and can improve dense matching results in regions, where photo-consistency computation is challenging due to low-textured areas, changing lighting conditions and reflections. A very interesting method has recently introduced by Ladicky *et al.* [Ladicky et al., 2010b], where the authors tried to estimate a pixel-wise classification and dense reconstruction jointly in a discrete optimization setting. They demonstrated attractive results on street-level images for a number of object classes, however, this concept might also work in the context of aerial imagery.

6.2.3 Virtual Cities and GIS Enrichment

While triangulation techniques offer an efficient representation for the visualization of uninterpreted large-scale 3D models, there is still a lack of ways to represent class-specific semantic information. Future work should therefore focus on the concept of a virtual city, where efficient data representations permit a thematic 3D visualization, but also the handling of search queries and user-specific data extraction. In particular, extracted data could then be used to refine and update GIS data in terms of accuracy, as well as to automatically increase the dimensionality of the existing systems from two to three dimensions.

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

Appendix A

Publications

My work at the Institute of Computer Graphics and Vision at Graz University of Technology led to the following list of published work. For the sake of completeness, these publications are reported in an inverse chronological order. Note that this thesis is mainly based on a subset of these papers (typed in bold letters).

(1) **Kluckner, S., Pock, T., Bischof, H.: Exploiting Redundancy for Aerial Image Fusion using Convex Optimization. In: Proceedings German Association for Pattern Recognition. (2010)**

(2) Mauthner, T., Kluckner, S., Roth, P.M., Bischof, H.: Efficient Object Detection Using Orthogonal NMF Descriptor Hierarchies. In: Proceedings German Association for Pattern Recognition. (2010)

(3) **Leberl, F., Bischof, H., Pock, T., Irschara, A., Kluckner, S.: Aerial Computer Vision for a 3D Virtual Habitat. IEEE Computer 43(6), pp 24-31 (2010)**

(4) **Kluckner, S., Bischof, H.: Large-Scale Aerial Image Interpretation Using A Redundant Semantic Classification. In: Proceedings International Society for Photogrammetry and Remote Sensing Symposium, Photogrammetric Computer Vision and Image Analysis. (2010)**

(5) **Kluckner, S., Bischof, H.: Image-based Building Classification and 3D Modeling with Super- Pixels. In: Proceedings International Society for Photogrammetry and Remote Sensing Symposium, Photogrammetric Computer Vision and Image Analysis. Best Paper Award. (2010)**

(6) **Kluckner, S., Unger, M., Pock, T., Bischof, H.: Large-scale Semantic Classification and Ortho-Image Fusion. In: BMVA Meeting, Aerial Image Analysis and Classification, Scientific Abstract. (2010)**

(7) Nguyen, T.T., Kluckner, S., Bischof, H., Leberl, F.: Aerial Photo Building Classification by Stacking Appearance and Elevation Measurements. In: Proceedings International Society for Photogrammetry and Remote Sensing Symposium, 100 Years ISPRS - Advancing Remote Sensing Science. (2010)

(8) Donoser, M., Kluckner, S., Bischof, H.: Object Tracking by Structure Tensor Analysis. In: Proceedings International Conference on Pattern Recognition. (2010)

(9) **Kluckner, S., Donoser, M., Bischof, H.: Super-Pixel Class Segmentation in Large-Scale Aerial Imagery. In: Proceedings Annual Workshop of the Austrian Association for Pattern Recognition. (2010)**

(10) **Kluckner, S., Bischof, H.: Semantic Classification by Covariance Descriptors Within a Randomized Forest. In: Proceedings International Conference on Computer Vision, Workshop on 3D Representation for Recognition (3dRR-09). (2009)**

(11) **Kluckner, S., Mauthner, T., Roth, P.M., Bischof, H.: Semantic Classification in Aerial Imagery by Integrating Appearance and Height Information. In: Proceedings Asian Conference on Computer Vision. (2009)**

(12) Leberl, F., Kluckner, S., Bischof, H.: Collection, Processing and Augmentation of VR Citites. In: Proceedings Photogrammetric Week, Stuttgart University. (2009)

(13) **Kluckner, S., Mauthner, T., Roth, P.M., Bischof, H.: Semantic Image Classification using Consistent Regions and Individual Context. In: Proceedings British Machine Vision Conference. (2009)**

(14) **Kluckner, S., Mauthner, T., Bischof, H.: A Covariance Approximation on Euclidean Space for Visual Tracking. In: Proceedings Annual Workshop of the Austrian Association for Pattern Recognition. (2009)**

(15) Kluckner, S., Bischof, H.: Pixel-wise Image Segmentation using Randomized Forest Classification and Shape Information. In: Proceedings Computer Vision Winter Workshop. (2009)

(16) **Kluckner, S., Pacher, G., Bischof, H., Leberl, F.: Objekterkennung in Luftbildern mit Methoden der Computer Vision durch kombinierte Verwendung von Redundanz, Farb- und Hoeheninformation. In: Proceedings Internationale 15. Geodaetische Woche. (2009)**

(17) Leberl, F., Kluckner, S., Pacher, G., Grabner, H., Bischof, H., Gruber, M.: Detecting Cars in Aerial Imagery for Improvements of Orthophotos and Digital Elevation Models. In: Proceedings ASPRS Annual Conference. (2008)

(18) Pacher, G., Kluckner, S., Bischof, H.: An Improved Car Detection using Street Layer Extraction. In: Proceedings Computer Vision Winter Workshop. (2008)

(19) Leberl, F., Bischof, H., Grabner, H., Kluckner, S.: Recognizing Cars in Aerial Imagery to Improve Orthophotos. In: Proceedings ACM International Symposium on Advances in Geographic Information Systems. (2007)

(20) Kluckner, S., Pacher, G., Grabner, H., Bischof, H., Bauer, J.: A 3D Teacher for Car Detection in Aerial Images. In: Proceedings International Conference on Computer Vision, Workshop on 3D Representation for Recognition (3dRR-07). (2007)

(21) Urschler, M., Kluckner, S., Bischof, H.: A Framework for Comparison and Evaluation of Nonlinear Intra-Subject Image Registration Algorithms. In: Proceedings International Conference on Medical Image Computing and Computer Assisted Intervention, ISC/NAMIC Workshop on Open Science. (2007)

Appendix B

Acronyms

CRF Conditional Markov random field

DSM Digital Surface Model

DTCWT Dual Tree Complex Wavelet Transform

DTM Digital Terrain Model

GIS Geographic Information System

GPU Graphics Processing Unit

GSD Ground Sampling Distance

KL Kullback-Leibler

LiDAR Light detection and ranging

LoD Level of Detail

PSNR Peak Signal-to-Noise Ratio

RF Randomized Forest

RMSE Root Mean Squared Error

SfM Structure from Motion

SVD Singular Value Decomposition

TV Total Variation

UT Unscented Transformation

List of Figures

List of Tables

Bibliography

[Agarwal et al., 2009] Agarwal, S., Snavely, N., Simon, I., Seitz, S. M., and Szeliski, R. (2009). Building Rome in a Day. In *Proceedings International Conference on Computer Vision*.

[Agarwala et al., 2006] Agarwala, A., Agrawala, M., Cohen, M., Salesin, D., and Szeliski, R. (2006). Photographing Long Scenes with Multi-viewpoint Panoramas. *ACM Transaction on Graphics (SIGGRAPH)*, 25.

[Amit et al., 1996] Amit, Y., Geman, A., and Geman, D. (1996). Shape Quantization and Recognition with Randomized Trees. *Neural Computation*, 9(7):1545–1588.

[Arsigny et al., 2007] Arsigny, V., Fillard, P., Pennec, X., and Ayache, N. (2007). Geometric Means in a Novel Vector Space Structure on Symmetric Positive-Definite Matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347.

[Baillard, 2008] Baillard, C. (2008). A Hybrid Method for Deriving DTMs from Urban DEMs. *International Archives of Photogrammetry and Remote Sensing*, XXXVII(3):109–111.

[Baillard and Zisserman, 2000] Baillard, C. and Zisserman, A. (2000). A Plane-Sweep Strategy For The 3D Reconstruction Of Buildings From Multiple Images. *International Archives of Photogrammetry and Remote Sensing*, XXXIII(2):56–62.

[Bay et al., 2006] Bay, H., Tuytelaars, T., and Van Gool, L. (2006). Surf: Speeded Up Robust Features. In *Proceedings European Conference on Computer Vision*.

[Bhattacharyya, 1945] Bhattacharyya, A. (1945). On a Measure of Divergence between two Statistical Populations Defined by their Probability Distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109.

[Bignone et al., 1996] Bignone, F., Henricsson, O., Fua, P., and Stricker, M. (1996). Automatic Extraction of Generic House Roofs from High Resolution Aerial Imagery. In *Proceedings European Conference on Computer Vision*, pages 83–96.

[Birchfield and Tomasi, 1998] Birchfield, S. and Tomasi, C. (1998). Depth Discontinuities by Pixel-to-Pixel Stereo. In *Proceedings International Conference on Computer Vision*.

[Boothby, 1975] Boothby, W. M. (1975). *Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press.

[Bosch et al., 2007] Bosch, A., Zisserman, A., and Munoz, X. (2007). Image Classification using Random Forests and Ferns. In *Proceedings International Conference on Computer Vision*.

[Boykov et al., 2001] Boykov, Y., Veksler, O., and Zabih, R. (2001). Efficient Approximate Energy Minimization via Graph Cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1222–1239.

[Bredies et al., 2010] Bredies, K., Kunisch, K., and Pock, T. (2010). Total Generalized Variation. *SIAM Journal on Imaging Sciences*, 3(3):492–526.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, pages 5–32.

[Brostow et al., 2008] Brostow, G. J., Shotton, J., Fauqueur, J., and Cipolla, R. (2008). Segmentation and Recognition Using Structure from Motion Point Clouds. In *Proceedings European Conference on Computer Vision*.

[Brown, 1976] Brown, D. C. (1976). The Bundle Adjustment - Progress and Prospects. *International Archives of Photogrammetry and Remote Sensing*, XXI(3).

[Candés et al., 2006] Candés, E., Laurent, D., Donoho, D., and Ying, L. (2006). Fast Discrete Curvelet Transforms. *Multiscale Modeling and Simulation*, 5(3):861–899.

[Carlavan et al., 2009] Carlavan, M., Weiss, P., Blanc-Féraud, L., and Zerubia, J. (2009). Complex Wavelet Regularization for Solving Inverse Problems in Remote Sensing. In *Proceedings Geoscience and Remote Sensing Society*.

[Chambolle and Pock, 2010] Chambolle, A. and Pock, T. (2010). A First-order Primal-dual Algorithm for Convex Problems with Applications to Imaging. Technical report, TU Graz.

[Champion and Boldo, 2006] Champion, N. and Boldo, D. (2006). A Robust Algorithm for Estimating Digital Terrain Models from Digital Surface Models in Dense Urban Areas. *International Archives of Photogrammetry and Remote Sensing*, XXXVI(3):1–6.

[Chan and Vese, 2001] Chan, T. and Vese, L. (2001). Active Contours Without Edges. *Transactions on Image Processing*, 10(2):266–277.

[Comaniciu and Meer, 2002] Comaniciu, D. and Meer, P. (2002). Mean Shift: A Robust Approach Toward Feature Space Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619.

[Cord et al., 1999] Cord, M., Jordan, M., Cocquerez, J.-P., and Paparoditis, N. (1999). Automatic Extraction and Modeling of Urban Buildings from High-resolution Aerial Images. *International Archives of Photogrammetry and Remote Sensing*, XXXII(3):187–192.

[Cox et al., 1996] Cox, I. J., Hingorani, S. L., Rao, S. B., and Maggs, B. M. (1996). A Maximum Likelihood Stereo Algorithm. *Computer Vision and Image Understanding*, 63(3):542–567.

[Criminisi et al., 2010] Criminisi, A., Shotton, J., Robertson, D., and Konukoglu, E. (2010). Regression Forests for Efficient Anatomy Detection and Localization in CT Studies. In *Medical Image Computing and Computer-Assisted Intervention*.

[Dalal and Triggs, 2005] Dalal, N. and Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Donoser et al., 2010] Donoser, M., Kluckner, S., and Bischof, H. (2010). Object Tracking by Structure Tensor Analysis. In *Proceedings International Conference on Pattern Recognition*.

[Dornaika and Bredif, 2008] Dornaika, F. and Bredif, M. (2008). An Efficient Approach to Building Superstructure Reconstruction Using Digital Elevation Maps. *International Archives of Photogrammetry and Remote Sensing*, XXXVII(3):179–185.

[Drauschke and Mayer, 2010] Drauschke, M. and Mayer, H. (2010). Evaluation of Texture Energies for Classification of Facade Images. *International Archives of Photogrammetry and Remote Sensing*, XXXVIII(3):257–262.

[Dryden et al., 2009] Dryden, I. L., Koloydenko, A., and Zhou, D. (2009). Non-Euclidean Statistics for Covariance Matrices, with Applications to Diffusion Tensor Imaging. *The Annals of Applied Statistics*, 3(3):1102–1123.

[Eckstein and Munkelt, 1995] Eckstein, W. and Munkelt, O. (1995). Extracting Objects From Digital Terrain Models. In *Remote Sensing and Reconstruction for Three-Dimensional Objects and Scenes, SPIE*, pages 43–51.

[Esser et al., 2009] Esser, E., Zhang, X., and Chan, T. (2009). A General Framework for a Class of First Order Primal-Dual Algorithms for TV Minimization. Technical Report 67, University of California, Los Angeles.

[Everingham et al., 2007] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2007). The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[Fadili et al., 2009] Fadili, M. J., Starck, J. L., and Murtagh, F. (2009). Inpainting and Zooming Using Sparse Representations. *Computer Journal*, 52(1).

[Felzenszwalb and Huttenlocher, 2004a] Felzenszwalb, P. F. and Huttenlocher, D. P. (2004a). Distance Transforms of Sampled Functions. Technical report, Cornell University, Computing and Information Science.

[Felzenszwalb and Huttenlocher, 2004b] Felzenszwalb, P. F. and Huttenlocher, D. P. (2004b). Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision*, 59(2):167–181.

[Fischler and Bolles, 1981] Fischler, M. A. and Bolles, R. C. (1981). Random Sample Consensus: A Paradigm for Model Fitting with Application to Image Analysis and Automated Cartography. *Communication Association and Computing Machine*, 24(6):381–395.

[Fitzgibbon et al., 2003] Fitzgibbon, A., Wexler, Y., and Zisserman, A. (2003). Image-based Rendering using Image-based Priors. In *Proceedings International Conference on Computer Vision*.

[Fletcher and Joshi, 2007] Fletcher, P. T. and Joshi, S. (2007). Riemannian Geometry for the Statistical Analysis of Diffusion Tensor Data. *Signal Processing*, 87(2):250–262.

[Förstner and Moonen, 1999] Förstner, W. and Moonen, B. (1999). A Metric for Covariance Matrices. Technical report, Department of Geodesy and Geoinformatics, Stuttgart University.

[Forsyth et al., 1996] Forsyth, D. A., Malik, J., Fleck, M. M., Greenspan, H., Leung, T., Belongie, S., Carson, C., and Bregler, C. (1996). Finding Pictures of Objects in Large Collections of Images. Technical report, UC Berkeley.

[Frahm et al., 2010] Frahm, J.-M., Georgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.-H., Enrique Dunn, B. C., Lazebnik, S., and Pollefeys, M. (2010). Building Rome on a Cloudless Day. In *Proceedings European Conference on Computer Vision*.

[Frey and Dueck, 2007] Frey, B. J. and Dueck, D. (2007). Clustering by Passing Messages Between Data Points. *Science*, 315:972–976.

[Fulkerson and Soatto, 2010] Fulkerson, B. and Soatto, S. (2010). Really Quick Shift: Image Segmentation on a GPU. In *Proceedings European Conference on Computer Vision, Workshop on Computer Vision using GPUs*.

[Fulkerson et al., 2009] Fulkerson, B., Vedaldi, A., and Soatto, S. (2009). Class Segmentation and Object Localization with Superpixel Neighborhoods. In *Proceedings International Conference on Computer Vision*.

[Gall and Lempitsky, 2009] Gall, J. and Lempitsky, V. (2009). Class-specific Hough Forests for Object Detection. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Galleguillos et al., 2008] Galleguillos, C., Babenko, B., Rabinovich, A., and Belongie, S. (2008). Weakly Supervised Object Recognition and Localization with Stable Segmentations. In *Proceedings European Conference on Computer Vision*.

[Gallup et al., 2007] Gallup, D., Frahm, J. M., Mordohai, P., Yang, Q., and Pollefeys, M. (2007). Real-time Plane-sweeping Stereo with Multiple Sweeping Directions. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Goesele et al., 2010] Goesele, M., Ackermann, J., Fuhrmann, S., Klowsky, R., Langguth, F., Muecke, P., and Ritz, M. (2010). Scene Reconstruction from Community Photo Collections. *Computer*, 43(6):48–53.

[Goldberger et al., 2003] Goldberger, J., Gordon, S., and Greenspan, H. (2003). An Efficient Image Similarity Measure based on Approximations of KL-Divergence between two Gaussian Mixtures. In *Proceedings International Conference on Computer Vision*.

[Goldluecke and Cremers, 2009] Goldluecke, B. and Cremers, D. (2009). A Superresolution Framework for High-Accuracy Multiview Reconstruction. In *Proceedings German Association for Pattern Recognition*.

[Gong et al., 2009] Gong, L., Wang, T., and Liu, F. (2009). Shape of Gaussians as Feature Descriptors. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Gould et al., 2009] Gould, S., Fulton, R., and Koller, D. (2009). Decomposing a Scene into Geometric and Semantically Consistent Regions. In *Proceedings International Conference on Computer Vision*.

[Gould et al., 2008] Gould, S., Rodgers, J., Cohen, D., Elidan, G., and Koller, D. (2008). Multi-class Segmentation with Relative Location Prior. *International Journal of Computer Vision*, 80(3):300–316.

[Grabner and Bischof, 2006] Grabner, H. and Bischof, H. (2006). On-line Boosting and Vision. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Grabner et al., 2008] Grabner, H., Nguyen, T., Grabner, B., and Bischof, H. (2008). On-line Boosting-based Car Detection from Aerial Images. *International Journal of Photogrammetry and Remote Sensing*, 63(3):382–396.

[Gruber, 2007] Gruber, M. (2007). UltraCamX, the New Digital Aerial Camera System by Microsoft Photogrammetry. In *Proceedings of the 50. Photogrammetric Week*.

[Gruber-Geymayer et al., 2005] Gruber-Geymayer, B. C., Klaus, A., and Karner, K. (2005). Data Fusion for Classification and Object Extraction. *International Archives of Photogrammetry and Remote Sensing*, XXXVI(3):125–130.

[Hartley and Zisserman, 2000] Hartley, R. I. and Zisserman, A. (2000). *Multiple View Geometry in Computer Vision*. Cambridge University Press.

[Hays and Efros, 2007] Hays, J. and Efros, A. A. (2007). Scene Completion Using Millions of Photographs. *ACM Transaction on Graphics (SIGGRAPH)*, 26.

[Heitz and Koller, 2008] Heitz, G. and Koller, D. (2008). Learning Spatial Context: Using Stuff to Find Things. In *Proceedings European Conference on Computer Vision*.

[Hirschmüller, 2006] Hirschmüller, H. (2006). Stereo Vision in Structured Environments by Consistent Semi-Global Matching. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Hirschmüller and Scharstein, 2009] Hirschmüller, H. and Scharstein, D. (2009). Evaluation of Stereo Matching Costs on Images with Radiometric Differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1582–1599.

[Hoiem et al., 2007] Hoiem, D., Stein, A., Efros, A. A., and Hebert, M. (2007). Recovering Occlusion Boundaries from a Single Image. In *Proceedings International Conference on Computer Vision*.

[Huber, 1981] Huber, P. (1981). *Robust Statistics*. Wiley, New York.

[Irschara, 2011] Irschara, A. (2011). *Image Based 3D Reconstruction for a Virtual Habitat*. PhD thesis, Graz University of Technology.

[Julier and Uhlmann, 1996] Julier, S. and Uhlmann, J. K. (1996). A General Method for Approximating Nonlinear Transformations of Probability Distributions. Technical report, Robotics Research Group, Department of Engineering Science, University of Oxford.

[Julier and Uhlmann, 1997] Julier, S. and Uhlmann, J. K. (1997). A New Extension of the Kalman Filter to Nonlinear Systems. In *International Symposium of Aerospace/Defense Sensing, Simulations and Controls*.

[Klaus et al., 2006] Klaus, A., Sormann, M., and Karner, K. (2006). Segment-Based Stereo Matching Using Belief Propagation and a Self-Adapting Dissimilarity Measure. In *Proceedings International Conference on Pattern Recognition*.

[Kluckner and Bischof, 2009] Kluckner, S. and Bischof, H. (2009). Semantic Classification by Covariance Descriptors Within a Randomized Forest. In *Proceedings International Conference on Computer Vision, Workshop on 3D Representation for Recognition (3dRR-09)*.

[Kluckner et al., 2009a] Kluckner, S., Mauthner, T., and Bischof, H. (2009a). A Covariance Approximation on Euclidean Space for Visual Tracking. In *Proceedings Annual Workshop of the Austrian Association for Pattern Recognition*.

[Kluckner et al., 2009b] Kluckner, S., Mauthner, T., Roth, P. M., and Bischof, H. (2009b). Semantic Image Classification using Consistent Regions and Individual Context. In *Proceedings British Machine Vision Conference*.

[Kohli et al., 2008] Kohli, P., Ladicky, L., and Torr, P. (2008). Robust Higher Order Potentials for Enforcing Label Consistency. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Kolbe et al., 2005] Kolbe, T., Gröger, G., and Plümer, L. (2005). CityGML - Interoperable Access to 3D City Models. In *Proceedings of the 1st International Symposium on Geo-information for Disaster Management*.

[Komodakis and Tziritas, 2007] Komodakis, N. and Tziritas, G. (2007). Approximate Labeling via Graph Cuts Based on Linear Programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1436–1453.

[Korč and Förstner, 2009] Korč, F. and Förstner, W. (2009). eTRIMS Image Database for Interpreting Images of Man-Made Scenes. Technical Report TR-IGG-P-2009-01, Department of Photogrammetry, University of Bonn.

[Kraus, 2004] Kraus, K. (2004). *Photogrammetrie: Geometrische Informationen aus Photographien und Laserscanner-Aufnahmen: Bd 1*. Gruyter, ISBN: 3110177080.

[Kullback and Leibler, 1951] Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

[Ladicky et al., 2010a] Ladicky, L., Sturgess, P., Alahari, K., Russell, C., and Torr, P. H. (2010a). What, Where And How Many? Combining Object Detectors and CRFs. In *Proceedings European Conference on Computer Vision*.

[Ladicky et al., 2010b] Ladicky, L., Sturgess, P., Russell, C., Sengupta, S., Bastanlar, Y., Clocksin, W., and Torr, P. H. (2010b). Joint Optimisation for Object Class Segmentation and Dense Stereo Reconstruction. In *Proceedings British Machine Vision Conference*.

[Ladstätter and Gruber, 2008] Ladstätter, R. and Gruber, M. (2008). Geometric Aspects Concerning the Photogrammetric Workflow of the Digital Aerial Camera UltraCamX. *International Archives of Photogrammetry and Remote Sensing*, XXXVII(1):521–525.

[Lafarge et al., 2008] Lafarge, F., Descombes, X., Zerubia, J., and Pierrot-Deseilligny, M. (2008). Automatic Building Extraction from DEMs using an Object Approach and Application to the 3D-city Modeling. *International Journal of Photogrammetry and Remote Sensing*, 63(3):365–381.

[Lafarge et al., 2010] Lafarge, F., Descombes, X., Zerubia, J., and Pierrot-Deseilligny, M. (2010). Structural Approach for Building Reconstruction from a Single DSM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):135–147.

[Lazebnik and Raginsky, 2009] Lazebnik, S. and Raginsky, M. (2009). An Empirical Bayes Approach to Contextual Region Classification. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Leberl et al., 2003] Leberl, F., Gruber, M., Ponticelli, M., Bernoegger, S., and Perko, R. (2003). The Ultracam Large Format Aerial Digital Camera System. In *Proceedings of the ASPRS Annual Convention*.

[Leibe et al., 2007] Leibe, B., Cornelis, N., Cornelis, K., and Van Gool, L. (2007). Dynamic 3D Scene Analysis from a Moving Vehicle. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Lepetit and Fua, 2006] Lepetit, V. and Fua, P. (2006). Keypoint Recognition using Randomized Trees. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Levi and Weiss, 2004] Levi, K. and Weiss, Y. (2004). Learning Object Detection from a Small Number of Examples: The Importance of Good Features. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Levinshtein et al., 2009] Levinshtein, A., Stere, A., Kutulakos, K. N., Fleet, D. J., Dickinson, S. J., and Siddiqi, K. (2009). TurboPixels: Fast Superpixels Using Geometric Flows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2290–2297.

[Li et al., 2008] Li, X., Hu, W., Zhang, Z., Zhang, X., Zhu, M., and Cheng, J. (2008). Visual Tracking via Incremental Log-Euclidean Riemannian Subspace Learning. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Liu et al., 2010] Liu, B., Gould, S., and Koller, D. (2010). Single Image Depth Estimation from Predicted Semantic Labels. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Lowe, 2004] Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110.

[Malisiewicz and Efros, 2007] Malisiewicz, T. and Efros, A. (2007). Improving Spatial Support for Objects via Multiple Segmentations. In *Proceedings British Machine Vision Conference*.

[Marszalek and Schmid, 2007] Marszalek, M. and Schmid, C. (2007). Accurate Object Localization with Shape Masks. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Matei et al., 2008] Matei, B. C., Sawhney, H. S., Samarasekera, S., Kim, J., and Kumar, R. (2008). Building segmentation for Densely Built Urban Regions using Aerial LIDAR Data. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Matikainen et al., 2007] Matikainen, L., Kaartinen, K., and Hyyppä (2007). Classification Tree Based Building Detection from Laser Scanner and Aerial Image Data. *International Archives of Photogrammetry and Remote Sensing*, XXXVI(3):280–287.

[Mauthner et al., 2010] Mauthner, T., Kluckner, S., Roth, P. M., and Bischof, H. (2010). Efficient Object Detection using Orthogonal NMF Descriptor Hierarchies. In *Proceedings German Association for Pattern Recognition*.

[Mayer, 1999] Mayer, H. (1999). Automatic Object Extraction from Aerial Imagery - A Survey Focusing on Buildings. *Computer Vision and Image Understanding*, 74(2):138–149.

[Moosmann et al., 2006] Moosmann, F., Triggs, B., and Jurie, F. (2006). Fast Discriminative Visual Codebooks using Randomized Clustering Forests. In *Advances in NIPS*.

[Nemirovski, 2004] Nemirovski, A. (2004). Prox-method with Rate of Convergence $O(1/t)$ for Variational Inequalities with Lipschitz Continuous Monotone Operators and Smooth Convex-concave Saddle Point Problems. *Journal on Optimization*, 15(1):229–251.

[Nesterov, 2005] Nesterov, Y. (2005). Smooth Minimization of Nonsmooth Functions. *Mathematical programming Series A*, 103:127–152.

[Nguyen et al., 2010] Nguyen, T. T., Kluckner, S., Bischof, H., and Leberl, F. (2010). Aerial Photo Building Classification by Stacking Appearance and Elevation Measurements. *International Archives of Photogrammetry and Remote Sensing*, XXXVIII(7):169–174.

[Nicklas, 2007] Nicklas, D. (2007). Nexus – A Global, Active and 3D Augmented Reality Model. In *Proceedings of the 51st Photogrammetric Week*, pages 325–334.

[Nikolova, 2004] Nikolova, M. (2004). A Variational Approach to remove outliers and impulse noise. *Journal of Mathematical Imaging and Vision*, 20(1-2):99–120.

[Nikolova et al., 2006] Nikolova, M., Esedoglu, S., and Chan, T. (2006). Algorithms for Finding Global Minimizers of Image Segmentation and Denoising Models. *SIAM Journal of Applied Mathematics*, 66(5):1632–1648.

[Nowak et al., 2006] Nowak, E., Jurie, F., and Triggs, B. (2006). Sampling Strategies for Bag-of-Features Image Classification. In *Proceedings European Conference on Computer Vision*.

[Ojala et al., 2002] Ojala, T., Pietikäinen, M., and Mäenpää, T. (2002). Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987.

[Olsson et al., 2009] Olsson, C., Byröd, M., Overgaard, N. C., and Kahl, F. (2009). Extending Continuous Cuts: Anisotropic Metrics and Expansion Moves. In *Proceedings International Conference on Computer Vision*.

[Paisitkriangkrai et al., 2008] Paisitkriangkrai, S., Shen, C., and Zhang, J. (2008). Performance evaluation of local features in human classification and detection. *IET Computer Vision*, 2(4):236–246.

[Pajares and de la Cruz, 2004] Pajares, G. and de la Cruz, J. M. (2004). A Wavelet-based Image Fusion Tutorial. *Pattern Recognition*, 37(9):1855 – 1872.

[Pantofaru et al., 2008] Pantofaru, C., Schmid, C., and Hebert, M. (2008). Object Recognition by Integrating Multiple Image Segmentations. In *Proceedings European Conference on Computer Vision*.

[Parish and Müller, 2001] Parish, Y. I. H. and Müller, P. (2001). Procedural Modeling of Cities. In *Proceedings Annual Conference on Computer Graphics and Interactive Techniques*, volume 28, pages 301–308.

[Pennec et al., 2006] Pennec, X., Pennec, X., and Fillard, P. (2006). A Riemannian Framework for Tensor Computing. *International Journal of Computer Vision*, 66(1):41–66.

[Pock et al., 2009] Pock, T., Chambolle, A., Cremers, D., and Bischof, H. (2009). A Convex Relaxation Approach for Computing Minimal Partitions. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Pock et al., 2008] Pock, T., Schoenemann, T., Graber, G., Bischof, H., and Cremers, D. (2008). A Convex Formulation of Continuous Multi-Label Problems. In *Proceedings European Conference on Computer Vision*.

[Porikli, 2005] Porikli, F. (2005). Integral Histogram: A Fast Way to Extract Histograms in Cartesian Spaces. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Porikli et al., 2006] Porikli, F., Tuzel, O., and Meer, P. (2006). Covariance Tracking using Model Update Based on Lie Algebra. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Portilla and Simoncelli, 2000] Portilla, J. and Simoncelli, E. P. (2000). A Parametric Texture Model based on Joint Statistics of Complex Wavelet Coefficients. *International Journal of Computer Vision*, 40(1):49–71.

[Porway et al., 2010] Porway, J., Wang, Q., and Zhu, S. (2010). A Hierarchical and Contextual Model for Aerial Image Parsing. *International Journal of Computer Vision*, 88(2):254–283.

[Potts, 1952] Potts, R. B. (1952). Some Generalized Order-Disorder Transformations. *Proceedings of the Cambridge Philosophical Society*, 48:106–109.

[Poullis and You, 2009] Poullis, C. and You, S. (2009). Automatic Reconstruction of Cities from Remote Sensor Data. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Rabinovich et al., 2006] Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., and Belongie, S. (2006). Objects in Context. In *Proceedings International Conference on Computer Vision*.

[Ranganathan, 2009] Ranganathan, A. (2009). Semantic Scene Segmentation using Random Multinomial Logit. In *Proceedings British Machine Vision Conference*.

[Ren and Malik, 2003] Ren, X. and Malik, J. (2003). Learning a Classification Model for Segmentation. In *Proceedings International Conference on Computer Vision*.

[Rottensteiner et al., 2004] Rottensteiner, F., Trinder, J., Clode, S., Kubik, K., and Lovell, B. (2004). Building Detection by Dempster-Shafer Fusion of LIDAR Data and Multispectral Aerial Imagery. In *Proceedings International Conference on Pattern Recognition*.

[Rudin et al., 1992] Rudin, L., Osher, S. J., and Fatemi, E. (1992). Nonlinear Total Variation Based Noise Removal Algorithms. *Physica D.*, 60:259–268.

[Russell et al., 2006] Russell, B. C., Efros, A. A., Sivic, J., Freeman, W. T., and Zisserman, A. (2006). Using Multiple Segmentations to Discover Objects and their Extent in Image Collections. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Russell and Torralba, 2009] Russell, B. C. and Torralba, A. (2009). Building a Database of 3D Scenes from User Annotations. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Santner et al., 2010] Santner, J., Pock, T., and Bischof, H. (2010). Interactive Multi-Label Segmentation. In *Proceedings Asian Conference on Computer Vision*.

[Saxena et al., 2008] Saxena, A., Sun, M., and Ng, A. Y. (2008). Make3D: Learning 3D Scene Structure from a Single Still Image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):824–840.

[Scharstein and Szeliski, 2001] Scharstein, D. and Szeliski, R. (2001). A Taxonomy and Evaluation of Dense Two-Frame StereoCorrespondence Algorithms. Technical report, Microsoft Research Technical Report.

[Schroff et al., 2008] Schroff, F., Criminisi, A., and Zisserman, A. (2008). Object Class Segmentation using Random Forests. In *Proceedings British Machine Vision Conference*.

[Selesnick et al., 2005] Selesnick, I. W., Baraniuk, R. G., and Kingsbury, N. G. (2005). The Dual-Tree Complex Wavelet Transform. *Signal Processing Magazine*, 22(6):123–151.

[Sharp, 2008] Sharp, T. (2008). Implementing Decision Trees and Forests on a GPU. In *Proceedings European Conference on Computer Vision*.

[Shotton et al., 2008] Shotton, J., Johnson, M., and Cipolla, R. (2008). Semantic Texton Forests for Image Categorization and Segmentation. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Shotton et al., 2007] Shotton, J., Winn, J., Rother, C., and Criminisi, A. (2007). TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context. *International Journal of Computer Vision*, 81(1):2–23.

[Sithole and Vosselman, 2005] Sithole, G. and Vosselman, G. (2005). Filtering of Airborne Laser Point Scanner Data Based on Segmented Point Clouds. *International Archives of Photogrammetry and Remote Sensing*, XXXVI(3):66–71.

[Sohn and Dowman, 2002] Sohn, G. and Dowman, I. (2002). Terrain Surface Reconstruction by the Use of Tetrahedron Model with the MDL Criterion. *International Archives of Photogrammetry and Remote Sensing*, XXXIV(1):336–344.

[Sorokin and Forsyth, 2008] Sorokin, A. and Forsyth, D. (2008). Utility data annotation with Amazon Mechanical Turk. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition, Workshop on Internet Vision*.

[Starck et al., 2004] Starck, J.-L., Elad, M., and Donoho, D. (2004). Image Decomposition via the Combination of Sparse Representations and a Variational Approach. *Transactions on Image Processing*, 14:1570–1582.

[Strecha et al., 2008] Strecha, C., Van Gool, L., and Fua, P. (2008). A Generative Model for True Orthorectification. *International Archives of Photogrammetry and Remote Sensing*, XXXVII(3):303–308.

[Sturgess et al., 2009] Sturgess, P., Alahari, K., Ladicky, L., and Torr, P. (2009). Combining Appearance and Structure from Motion Features for Road Scene Understanding. In *Proceedings British Machine Vision Conference*.

[Taillandier, 2005] Taillandier, F. (2005). Automatic Building Reconstruction from Cadastral Maps and Aerial Images. *International Archives of Photogrammetry and Remote Sensing*, XXXVI(3):105–110.

[Tikhonov, 1943] Tikhonov, A. N. (1943). On the Stability of Inverse Problems. *Doklady Akademii Nauk SSSR 39*, 5:195–198.

[Toshev et al., 2010] Toshev, A., Mordohai, P., and Taskar, B. (2010). Detection and Parsing Architecture at City Scale from Range Data. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Tukey, 1974] Tukey, J. W. (1974). Mathematics and the Picturing of Data. In *Proceedings Int. Congress of Mathematics*, volume 2, pages 523–531.

[Tuzel et al., 2006] Tuzel, O., Porikli, F., and Meer, P. (2006). Region covariance: A Fast Descriptor for Detection and Classification. In *Proceedings European Conference on Computer Vision*.

[Tuzel et al., 2007] Tuzel, O., Porikli, F., and Meer, P. (2007). Human Detection via Classification on Riemannian Manifolds. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Tyagi and Davis, 2008] Tyagi, A. and Davis, J. W. (2008). A Recursive Filter for Linear Systems on Riemannian Manifolds. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Unger et al., 2010] Unger, M., Pock, T., Werlberger, M., and Bischof, H. (2010). A Convex Approach for Variational Super-Resolution. In *Proceedings German Association for Pattern Recognition*.

[Unnikrishnan et al., 2007] Unnikrishnan, R., Pantofaru, C., and Hebert, M. (2007). Toward Objective Evaluation of Image Segmentation Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):929–944.

[Vedaldi and Soatto, 2008] Vedaldi, A. and Soatto, S. (2008). Quick Shift and Kernel Methods for Mode Seeking. In *Proceedings European Conference on Computer Vision*.

[Verbeek and Triggs, 2007] Verbeek, J. and Triggs, B. (2007). Scene Segmentation with CRFs Learned from Partially Labeled Images. In *Advances in NIPS*.

[Viola and Jones, 2004] Viola, P. and Jones, M. (2004). Robust Real-time Object Detection. *International Journal of Computer Vision*, 57(2):137–154.

[Vosselman and Dijkman, 2001] Vosselman, G. and Dijkman, S. (2001). 3D Building Model Reconstruction from Point Clouds and Ground Plans. *International Archives of Photogrammetry and Remote Sensing*, XXXIV(3):37–44.

[Weidner and Förstner, 1995] Weidner, U. and Förstner, W. (1995). Towards Automatic Building Extraction from High-resolution Digital Elevation Models. *International Journal of Photogrammetry and Remote Sensing*, 50(4):38–49.

[Werner and Zisserman, 2002] Werner, T. and Zisserman, A. (2002). New Techniques for Automated Architectural Reconstruction from Photographs. In *Proceedings European Conference on Computer Vision*.

[Winn et al., 2005] Winn, J., Criminisi, A., and Minka, T. (2005). Object Categorization by Learned Universal Visual Dictionary. In *Proceedings International Conference on Computer Vision*.

[Woodford et al., 2007] Woodford, O. J., Reid, I. D., Torr, P. H., and Fitzgibbon, A. W. (2007). On New View Synthesis Using Multiview Stereo. In *Proceedings British Machine Vision Conference*.

[Wu and Nevatia, 2008] Wu, B. and Nevatia, R. (2008). Optimizing Discrimination-efficiency Tradeoff in Integrating Heterogeneous Local Features for Object Detection. In *Proceedings IEEE Conference Computer Vision and Pattern Recognition*.

[Xiao and Quan, 2009] Xiao, J. and Quan, L. (2009). Multiple View Semantic Segmentation for Street View Images. In *Proceedings International Conference on Computer Vision*.

[Yi et al., 2009] Yi, Z., Criminisi, A., Shotton, J., and Blake, A. (2009). Discriminative, Semantic Segmentation of Brain Tissue in MR Images. In *Medical Image Computing and Computer-Assisted Intervention*.

[Yoon and Kweon, 2006] Yoon, K. J. and Kweon, I. S. (2006). Adaptive Support-Weight Approach for Correspondence Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):650–656.

[Zach et al., 2007] Zach, C., Pock, T., and Bischof, H. (2007). A Globally Optimal Algorithm for Robust TV-L^1 Range Image Integration. In *Proceedings International Conference on Computer Vision*.

[Zebedin, 2010] Zebedin, L. (2010). *Automatic Reconstruction of Urban Environments from Aerial Images*. PhD thesis, Graz University of Technology.

[Zebedin et al., 2008] Zebedin, L., Bauer, J., Karner, K., and Bischof, H. (2008). Fusion of Feature- and Area-Based Information for Urban Buildings Modeling from Aerial Imagery. In *Proceedings European Conference on Computer Vision*.

[Zebedin et al., 2006] Zebedin, L., Klaus, A., Gruber-Geymayer, B., and Karner, K. (2006). Towards 3D Map Generation from Digital Aerial Images. *International Journal of Photogrammetry and Remote Sensing*, 60(6):413–427.