DGK Deutsche Geodätische Kommission
bei der Bayerischen Akademie der Wissenschaften

Reihe C Dissertationen Heft Nr. 669

Jan Dirk Wegner

Detection and height estimation of buildings from SAR and optical images using conditional random fields

München 2011

Verlag der Bayerischen Akademie der Wissenschaften
in Kommission beim Verlag C. H. Beck

ISSN 0065-5325 ISBN 978-3-7696-5081-5

Diese Arbeit ist gleichzeitig veröffentlicht in:
Wissenschaftliche Arbeiten der Fachrichtung Geodäsie und Geoinformatik der Leibniz Universität Hannover
ISSN 0174-1454, Nr. 295, Hannover 2011

DGK Deutsche Geodätische Kommission

bei der Bayerischen Akademie der Wissenschaften

Reihe C Dissertationen Heft Nr. 669

Detection and height estimation of buildings from SAR and optical images using conditional random fields

Von der Fakultät für Bauingenieurwesen und Geodäsie
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des Grades
Doktor-Ingenieur (Dr.-Ing.)
genehmigte Dissertation

von

Dipl.-Ing. Jan Dirk Wegner

geboren am 09.02.1982 in Oldenburg

München 2011

Verlag der Bayerischen Akademie der Wissenschaften
in Kommission bei der C. H. Beck'schen Verlagsbuchhandlung München

ISSN 0065-5325 ISBN 978-3-7696-5081-5

Diese Arbeit ist gleichzeitig veröffentlicht in:
Wissenschaftliche Arbeiten der Fachrichtung Geodäsie und Geoinformatik der Leibniz Universität Hannover
ISSN 0174-1454, Nr. 295, Hannover 2011

Adresse der Deutschen Geodätischen Kommission:

DGK

Deutsche Geodätische Kommission

Alfons-Goppel-Straße 11 ● D – 80 539 München

Telefon +49 – 89 – 23 031 1113 ● Telefax +49 – 89 – 23 031 - 1283 / - 1100
e-mail hornik@dgfi.badw.de ● http://www.dgk.badw.de

Prüfungskommission

Vorsitzender:	Prof. Dr.-Ing. Udo Nackenhorst
Referenten:	Prof. Dr.-Ing. Uwe Sörgel
	Prof. Dr.-Ing. Uwe Stilla
	Prof. Dr.-Ing. Monika Sester

Tag der Einreichung der Arbeit:	20.06.2011
Tag der mündlichen Prüfung:	02.08.2011

ISSN 0065-5325

ISBN 978-3-7696-5081-5

Erklärung

Ich erkläre, dass ich die vorliegende Dissertation selbständig verfasst habe, die benutzten Hilfsmittel vollständig angegeben sind und die Dissertation nicht als Diplomarbeit, Masterarbeit oder andere Prüfungsarbeit verwendet wurde. Weiterhin erkläre ich, dass ich keine anderen Promotionsgesuche eingereicht habe.

Hannover, 20. Juni 2011

Statement

I state that this dissertation has been written entirely by myself. No further sources besides the ones noted in the bibliography have been used and this dissertation has not been submitted as Diploma thesis, Master thesis or any other written examination. Furthermore I state that I have not applied for any other conferral of a doctorate.

Hannover, 20th June 2011

Abstract

Single buildings in urban scenes are visible in very high-resolution data of *synthetic aperture radar* (SAR) sensors like TerraSAR-X or Aes-1. All-weather and all-day data acquisition capability make SAR a valuable tool for rapid mapping in crisis situations, but geometric effects as layover and the narrow signal spectrum hamper automatic data analysis. Complementary information derived from a multi-spectral optical high-resolution image can ease interpretability, but both data cannot be fused pixel-wise due to three-dimensional effects calling for feature-based fusion.

Contextual information can significantly improve classification if features are insufficient to discriminate different object categories. Usually rule-based methods are used, needing manual parameter tuning anew for each scene. Novel methods have to be developed to detect building objects with a comprehensive contextual probabilistic approach, learning its parameters from training data in order to guarantee applicability to any scene.

Estimation of building heights is important to determine whether particular buildings are higher than an expected flooding level, for example. Current methods rely on only few single measurements or simulations without a sound stochastic interpretation thus struggling in terms of validity and reliability.

This thesis adresses four essential aspects: First, *appropriate features in SAR and optical data are extracted.* Second, *novel approaches to probabilistic formulation of urban scene context are introduced.* Third, *new techniques to measure building heights based on a combination of one SAR acquisition and an optical image are proposed.* Finally, *a rigorous stochastic approach is suggested to derive a single robust height per building with a corresponding precision.*

Novel object-context formulations within a Conditional Random Field (CRF) framework are introduced. A graph is set up on image regions generated by a segmentation, better preserving object boundaries than standard patch grids. Its irregular structure, representing the scene topology, is exploited for contextual learning and object detection. A novel way to capture patterns in partially labeled data, so-called *implicit scene context* (ISC), is proposed. Concerning building height estimation, measures relying on combined SAR and optical observations are developed. Multiple heights per building are combined in a stringent stochastic framework based on *least squares adjustment with functionally dependent parameters.* It allows to assess height accuracies that can be achieved under optimal conditions.

Experiments with one SAR acquisition and an optical image reveal potentials and limitations of the proposed methods. The developed CRF approaches can easily be transferred to different scenes and to entirely different data overcoming characteristic drawbacks of rule-based or only partially probabilistic methods. Building detection results are very promising, but unveil need for, first, more sophisticated features, second, an even higher level of detail concerning context formulation within the CRF. Least squares adjustment proves to provide robust building heights, enabling the assessment of their validity and reliability through precision values. Height estimation with meter accuracy is possible.

Keywords: random fields, contextual classification, probabilistic modelling, fusion, building detection, graph, height estimation, synthetic aperture radar (SAR), high resolution, least squares adjustment, feature extraction

Zusammenfassung

In sehr hoch aufgelösten SAR-Daten können einzelne städtische Gebäude erkannt werden. Geometrische Effekte und schmales Signalspektrum erschweren jedoch die automatisierte Datenanalyse. Komplementäre Informationen aus multispektralen optischen Fernerkundungsbildern können die Objektdetektion entscheidend verbessern. Bedingt durch dreidimensionale Effekte, insbesondere in städtischen Gebieten, können SAR-Daten und optische Bilder allerdings nicht direkt pixel-basiert fusioniert werden. Deshalb wird in der Dissertation eine kombinierte Auswertung beider Datentypen auf Merkmals-Ebene durchgeführt.

Genügen direkte Merkmale nicht zur Objektdetektion, kann Kontextwissen diese entscheidend verbessern. Aktuelle Ansätze basieren häufig auf einer großen Anzahl Regeln, deren Parameter für jede Szene manuell eingestellt werden. Die Entwicklung neuer kontext-basierter probabilistischer Ansätze, die ihre Parameter auf Grundlage von Trainingsdaten erlernen, ermöglicht eine automatische Anpassung an neue Szenen.

Nach der erfolgreichen Detektion von Gebäuden kann es für eine Vielzahl von Anwendungen, z.B. während einer Flutkatastrophe, wünschenswert sein, eine Höhe pro Gebäude zu schätzen. Aktuelle Ansätze zur Bestimmung von Gebäudehöhen nutzen oft nur einzelne Möglichkeiten der Höhenmessung. Zudem sind diese in der Regel nicht stochastisch interpretierbar, essentielle Aussagen zu Genauigkeit und Zuverlässigkeit können nicht getroffen werden.

Diese Doktorarbeit behandelt vier wesentliche Aspekte: 1) Die Extraktion von *Gebäudemerkmalen in SAR-Daten und optischen Bildern*, 2) neue Ansätze zur *probabilistischen Formulierung von urbanem Szenenkontext*, 3) innovative Methoden zur *Höhenbestimmung von Gebäuden* mittels *Kombination eines SAR-Datensatzes und eines optischen Bildes*, 4) einen *stochastisch strengen Ansatz zur Schätzung einer einzigen Höhe pro Gebäude* aus mehreren gemessenen, der jeweils eine *Genauigkeitsaussage* trifft.

Neue Möglichkeiten der Formulierung von Kontextwissen basierend auf Conditional Random Fields (CRF) werden eingeführt. Unregelmäßige Graphstrukturen von Bildregionen, die die Szenentopologie repräsentieren, ersetzen Gitter quadratischer Bildteilflächen. Diese Graphen werden zum Erlernen von Objektkontext und zur anschließenden Objektdetektion genutzt. Des Weiteren wird ein neuer Ansatz zum Erlernen von Kontext in nur teilweise semantisch belegten Trainingsdaten eingeführt. Neue Möglichkeiten der Bestimmung von Gebäudehöhen werden beschrieben und ein Gauß-Helmert-Model eingeführt, das alle Messungen pro Gebäude ausgleicht und mit einer Standardabweichung versieht.

Experimente mit einem SAR-Datensatz und einem optischen Bild lassen sowohl Vorteile als auch Einschränkungen der vorgeschlagenen Methoden erkennen. Die entwickelten lernenden CRF-Ansätze können ohne Änderungen direkt auf andere Datensätze angewendet werden, ein großer Vorteil gegenüber regelbasierten Techniken. Die Gebäudedetektionsergebnisse sind sehr vielversprechend, jedoch bieten sich spezifischere Merkmale sowie eine noch komplexere Modellierung kontextueller Objektrelationen zur weiteren Ergebnisverbesserung an. Die Gebäudehöhenmessungen kombiniert im Gauß-Helmert-Model liefern Ergebnisse mit Genauigkeiten im Meterbereich.

Schlagworte: Zufallsfelder, kontext-basierte Klassifizierung, probabilistische Modellierung, Fusion, Gebäudedetektion, Graph, Höhenschätzung, Radar mit synthetischer Apertur (SAR), hohe Auflösung, Ausgleichung, Merkmalsextraktion

Table of Symbols

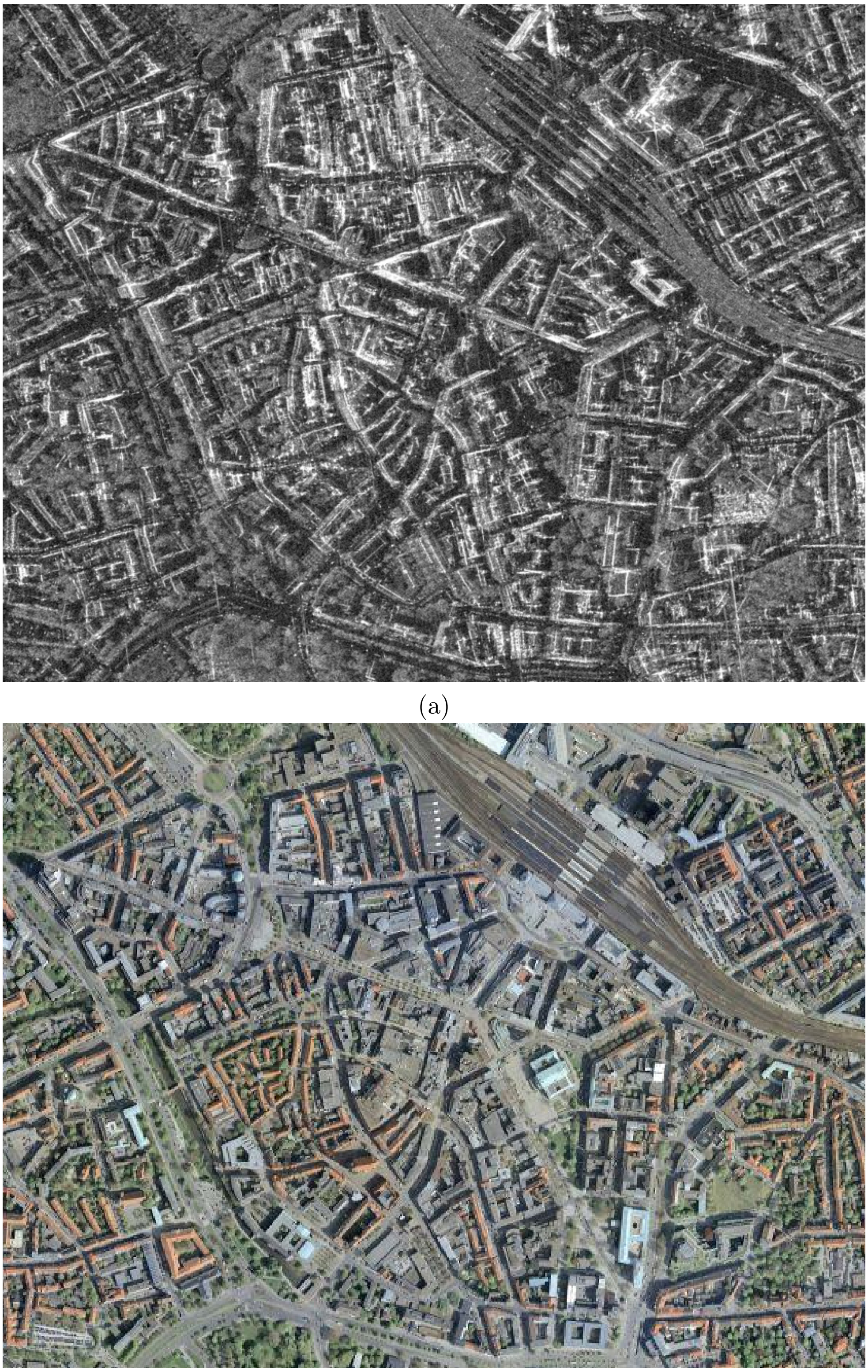
Symbol	Meaning
Probabilistic modelling	
$P(x)$	marginal probability of data x
$P(y, x)$	joint probability of data x and label y
$P(y\|x)$	conditional probability of y conditioned on x
$A_i(\mathbf{x}, y_i)$	association potential of node i
$I_{ij}(\mathbf{x}, y_i, y_j)$	interaction potential of considering nodes i and j
$Z(\mathbf{x})$	partition function
$\mathbf{h}_i(\mathbf{x})$	node feature vector (with weights $\mathbf{w}$ to be trained)
$\boldsymbol{\mu}_{ij}(\mathbf{x})$	edge feature vector (with weights $\mathbf{v}$ to be trained)
Building height estimation	
h_s	height via sun shadow
h_{pd}	height via optical perspective distortion
h_{db}	height via overlap of roof edge and double-bounce line
h_{InSAR}	robust maximum InSAR height in layover ramp
h_l	height via layover in SAR magnitude image
$h_{b,noI}$	adjusted building height excluding h_{InSAR}
h_b	adjusted height combining all available height measurements
h_L	reference height of airborne laserscanning (LiDAR)
$\hat{\sigma}_b$	posterior standard deviation after height adjustment
$\Delta_{b,L}$	difference of adjusted height h_b to LiDAR reference height h_L
Least squares adjustment	
$\hat{\boldsymbol{l}}$	adjusted observations
$\hat{\boldsymbol{x}}$	adjusted height corrections to h_0
$\mathbf{B}$	first partial derivatives with respect to observations in $\mathbf{l}$
$\mathbf{A}$	first partial derivatives with respect to parameters in $\mathbf{x}$
$\mathbf{v}$	difference between original and adjusted observations
$\mathbf{Q}_{ll}$	variance-covariance matrix
$\mathbf{P}$	weight matrix

Contents

1. Introduction

Synthetic aperture radar (SAR) has become a very important remote sensing technique in the last two decades. Two key features of SAR in comparison to optical sensors are that it is independent of daylight and its all-weather data acquisition capability. Reasons are the longer signal wavelength (usually 3 to 25 centimeters) compared to the visible spectrum and the active sensor principle. Operating spaceborne systems like ERS-2 and ENVISAT provide rather coarse spatial resolutions (e.g., 25m ground sampling distance). Information extraction from those images is often restricted to radiometric properties; a typical application is land cover classification. Structures of settlement areas can usually be characterized only in a rather generalized manner, inner city areas and suburbs may be distinguished. In SAR data of one meter geometric resolution collected by modern spaceborne sensors like TerraSAR-X and Cosmo-SkyMed, the geometric extent of individual objects like bridges, buildings, and roads is visible. In figure 1.1(a) a TerraSAR-X high-resolution spotlight image of the city center of Hannover, Germany, is shown. Objects and object parts are visible in very high-resolution data of approximately one meter. We can recognize buildings, vegetated areas, and the railroad tracks in the upper right corner of the image.

Airborne sensors image the urban scene with even more detail. However, shadowing and layover effects, typical for SAR image acquisitions in urban areas, always complicate interpretation. Small buildings are often occluded by higher ones and facades overlap with trees and cars on the streets. In addition, the appearance of a building in the image highly depends on the sensor's aspect. We can thus add data from another sensor to complement SAR data. Optical images have the advantage of being widely available. In addition, they can provide complimentary information about objects on the ground because optical sensors differ from SAR sensors in terms of geometry and radiometry (cf. 1.1(a) and (b)). Optical sensors are passive sensors performing angular measurements, whereas SAR sensors actively emit pulses and measure distances towards the objects. SAR sensors have a very high dynamic range of radiometric values, but their signal is limited to a small spectrum in the microwave domain. Optical sensors feature a lower dynamic range, but are capable of recording multi-spectral information of the sunlight reflected at an object. Therefore, a combination of optical and SAR data is able to provide a much richer description of an object on the ground then one single data source. It is particularly convenient in highly complex scenes (like shown in Fig. 1.1) containing a great amount of different object categories. In urban scenes we face the challenge of discriminating buildings from various other categories like streets, vegetated areas, and parking lots. Later on we will see in section 2.1.4 which hints in SAR and optical data support building detection.



(a)

(b)

Figure 1.1.: (a) TerraSAR-X high-resolution spotlight image (range direction left to right) of the city Hannover, Germany (©DLR), (b) corresponding aerial photo (©Google)

In addition to complimentary data of two different sensor types we can support building detection in complex urban scenes through the exploitation of object-context. A building is not only described by a certain roof color and texture, but also by contextual attributes in the local vicinity. For example, sun shadow is a good hint to a three-dimensional object, front yards often occur at buildings, and driveways lead towards them. This is what we mean by local context of a building object. If we enlarge our view spatially, buildings in urban areas are often aligned with streets, pavements are located in-between street and buildings. This relationship of different categories of objects or object parts is what we call regional context. One further extension is global urban context which may encode that small gable roof buildings are likely to occur in suburban areas, whereas high-rise buildings are often located in the city centers. *In this thesis different possibilities to incorporate and learn context in a probabilistic approach are proposed with focus on local and regional urban context.*

Furthermore, the fusion of optical and SAR data also provides new means for building height estimation. Different viewing geometries of the sensors enable several new ways to measure building heights. First, geometric effects like layover in SAR data and effects caused by the central perspective of an optical camera contain height information. Heights that are separately measured in optical and SAR data can thus be combined to estimate one single robust height per building. Second, we can also directly combine both data to measure a building's height. All obtained heights of one building, separate and combined ones, have to be jointly evaluated in order to achieve a final robust building height estimate. The weighting of each possible height measurement should have an influence on the final height depending on its accuracy. A precision measure should be assigned to each final height in order to evaluate its overall quality. We achieve the goals aforementioned by *introducing least squares adjustment, a stochastically sound approach, to building height estimation based on one SAR acquisition and an optical image.* In the following section the proposed approaches are motivated and the main contributions of this thesis are clearly stated.

1.1. Motivation and objectives

In this thesis focus is on SAR data and optical imagery of urban areas. Cities are of particular interest because they are densely inhabited by humans, any change may immediately affect lifes. It is essential for human societies to monitor and map ongoing activities in those densely populated areas. One major way to meet this requirement is to use remote sensing as a primary source of information. Such data gains particular importance in crisis situations (e.g., natural disasters) because large areas can be mapped within a relatively short time. Rapid mapping is needed for instant response actions of the public authorities and aid agencies. Due to the immediate need of post-crisis information it is often impossible to acquire rich and comprehensive data that would originally be used for urban scene analysis (e.g., airborne laserscanning or optical stereo imagery). SAR sensors are the appropriate choice for rapid mapping due to their all-weather and all-day capabilities. SAR data can immediately be acquired after the disaster by a high-resolution SAR sensor passing once over the scene. This SAR sensor can either be mounted on an aircraft (airborne) or on a satellite platform (spaceborne). Often, an optical image acquired before the disaster is available, too. We now face the challenge of automatic scene analysis based on merely one SAR acquisition and an optical image. The most important objects in urban areas are buildings, thus we focus on building detection.

The first objective of this thesis is to develop an innovative solution for the detection of buildings in urban areas merging information derived from one high-resolution SAR acquisition and one optical image. One SAR acquisition can either be one single SAR image or an interferometric SAR image pair acquired in single-pass mode with a certain baseline. At this point it should be noted that we do not want to perform change detection. The aim is to investigate joint use of complementary data of those two different sensor types for building detection. In case local evidence about a certain building is sparse, knowledge about the typical structure of the scene can support object detection. This contextual information reduces the number of possible locations and features to be considered.

The majority of object detection approaches incorporating context information relies on model knowledge translated to a set of rules. A model of an object that is to be detected can be formulated either implicitly or explicitly. Implicit model representation often interweaves model knowledge with design and work-flow of data processing, which can become inflexible if dealing with a new object category. Approaches using explicit object models are called knowledge-based approaches (e.g., production nets or semantic nets). Sets of rules explicitly formulate the precise model of an object (and its context) independent of data processing (e.g., [Stilla, 1995; Koch et al., 1997; Kunz et al., 1997; Soergel et al., 2003b]). Advantages are that prior expert knowledge can directly be modelled and graphical representations of object relations can be intuitively understood. Furthermore, knowledge-based systems provide more flexibility compared to systems modelling objects implicitly because only the explicit object model has to be adapted for a new object category without changing the entire processing chain. Additional possibilities for object detection besides production nets and semantic nets are fuzzy logic [Zadeh, 1965] (remote sensing applications, e.g., [Benz et al., 2004; Tóvári & Vögtle, 2004]) and Dempster-Shafer evidential theory [Shafer, 1976] (remote sensing applications, e.g., [Quint & Sties, 1996; Hégarat-Mascle et al., 1997; Rottensteiner et al., 2007; Poulain et al., 2011]). They also formulate object model knowledge rather intuitively and results can well be understood by human interpreters. In addition, Dempster-Shafer approaches provide the possibility of modelling uncertainty explicitly.

Some principle drawbacks of the aforementioned approaches exist. A first one is that usually lots of different parameters have to be set anew for each scene. Moreover, they cannot cope with information that has not been explicitly modelled beforehand. In case of highly complex scenarios, as urban areas, human experts may not be able to recognize all underlying rules. From a classification perspective we can view such a complex problem as a very high-dimensional feature space, where distinctive feature distributions are to be found. In a rule-based system, an expert would have to assess the importance of each distribution for discriminating classes of interest manually. A weight would have to be assigned to each feature and, most notably, to its combination with all other features. Humans are able to do this manually for distributions of single features. Discriminative joint distributions of two or three features may still be recognized, but beyond three features we can hardly tell the exact weights because the dimension of the problem gets too high. If distinctive patterns of very high dimension exist in feature space discriminating the desired object categories, human experts will potentially not recognize them leading to missing rules in the model. These underlying patterns in high-dimensional joint distributions can be captured via computer-based learning techniques (i.e., machine learning). In addition, learning procedures make classification approaches adaptable to scenes of a new environment by re-adjusting weights of features. Reconsidering rule-based methods, machine learning is often not integrated or only for a small percentage of the entire parameter set. Due to being tailored to one specific task (e.g., building detection in remote sensing data) those approaches can hardly be transfered to different tasks or scenarios (e.g., building facade detection in terrestrial images) without an expert rearranging or defining new rules. However, rule-based approaches can be reformulated in a comprehensive probabilistic way as directed graphical models, so-called Bayesian networks, as done by Stilla & Hedman [2010], for example.

In this thesis it is proposed to choose a *contextual probabilistic approach learning its parameters from a database of labeled training data.* A family of methods capable of meeting all requirements are graphical models, more precisely Conditional Random Fields (CRF).

The second objective is the accuracy assessment of building height estimation based on a single SAR acquisition and one optical image. Considering the crisis scenario, building heights might be benefitial too, for example if the region risks to be flooded. Several works have already dealt with height measurements based merely on SAR data or a combination with optical data (details in section 1.3.3). However, none fully exploits all different height measurement possibilities that arise if dealing with one SAR acquisition and an optical image. Most of them only handle flat roof buildings and do not consider gable roof buildings. Furthermore, heights have not been determined within a sound stochastic approach that combines different height measurements to provide one final robust height estimate. In addition, the accuracy of the building heights that can theoretically be achieved has not been investigated, yet. Therefore, we first need to introduce additional ways to measure building heights based on a combination of the given data. Second, we have to design a stochastic approach that weights the influence of each single height measurments according to its accuracy. It should also assign a precision value to the final height of each flat roof and gable roof building.

These requirements will be met by introducing *new ways of measuring building heights combining SAR data and optical image* and by *evaluating all single heights within a least squares adjustment approach.*

Summarizing the goals of this thesis in one sentence: ***The aim is to automatically detect buildings based on features of one high-resolution SAR acquisition and one optical image, to integrate contextual information into a probabilistic framework, and to estimate the building heights.***

1.2. Reader's guide

This thesis is structured as follows. First, state-of-the-art approaches dealing with fusion of optical and SAR data, context-based classification, and building height estimation are reviewed. In Chapter 2 fundamentals of two major topics are described: the sensors' characteristics and probabilistic modelling. First, differences of SAR sensors and optical sensors are explained with emphasis on the appearance of buildings. Second, the reader is familiarized with basic concepts of probabilistic models for classification with particular focus on context-based methods. Additionally, an insight into training and inference is provided in Appendix A. The methodology of the developed approaches is explained in detail in Chapter 3. Different ways to formulate contextual knowledge within the framework of Conditional Random Fields in order to detect buildings are shown. Then, a least squares approach to building height estimation based on one SAR acquisition and an optical image is presented. In the following Chapter 4 previously introduced methods are applied to test data and results are presented. Those results will be discussed and evaluated in Chapter 5. Finally, conclusions are drawn and directions for future research proposed.

1.3. State-of-the-art

In order to resolve the task presented in section 1.1, a variety of scientific research areas is touched. This section presents the current state-of-the-art of the three most important fields of research regarding this project: *Fusion of optical and SAR data* (1.3.1), *classification using context* (1.3.2), and *building height estimation* (1.3.3). Methods of the first and the third research area have been proposed by scientists belonging to the remote sensing community. We should reconsider at this point that the given data are limited to only one single SAR acquisition and an optical image. In subsections 1.3.1 and 1.3.3 focus is on approaches with a similar configuration.

Major research of the second topic (1.3.2) has been done in the computer vision and machine learning community. In this thesis their findings are introduced to the remote sensing community and extended. Today, most object detection approaches in remote sensing directly formulate model knowledge in a non-probabilistic way, usually without a learning step (i.e., all parameters have to be adjusted manually). The aim is to avoid this direct formulation of rules based on a very specific object model. We need to learn object appearances within one concise and comprehensive probabilistic framework instead. Furthermore, we want to learn the context of an object, which is the typical environment of the object. In case a new scene arrives that is not contained in the database we can simply add it to the training database and retrain the parameters.

In the long term we will achieve a fully automated procedure, the main objective of object classification in remote sensing. In order to familiarize the reader with a remote sensing background with contextual probabilistic object classification, a comprehensive overview of recent developments in the computer vision and machine learning communities is provided in subsection 1.3.2.

1.3.1. Fusion of optical and SAR data

We have to define the term fusion first because it is used with different meanings in the remote sensing community. Fusion can imply four different ways of data processing that have to be carefully distinguished:

- *automatic co-registration* of data acquired by SAR and optical sensors [Toutin, 1995; Dare & Dowman, 2000; Inglada & Giros, 2004; Hong & Schowengerdt, 2005; Wegner, 2007; Suri et al., 2009; Suri & Reinartz, 2010],
- *pixel-based fusion* of grey-values of SAR data and an optical image with the primary aim of an improved visualization [Ehlers & Tomowski, 2008; Soergel et al., 2008],
- *feature-based fusion* derived from data acquired by SAR and optical sensors with the goal of segmentation [Lombardo et al., 2003], land cover classification [Schistad et al., 1996; Macri-Pellizzeri et al., 2002; Hégarat-Mascle et al., 1997; Waske & Benediktsson, 2007] or object detection,
- *decision-based fusion* of different classification achieved with data of different sensor types as input [Benediktsson et al., 1990; Serpico & Roli, 1995; Briem et al., 2002; Waske & van der Linden, 2008].

The understanding of fusion in this work is following the third category and focus particularly is on object detection. Features of optical and SAR data are combined in order to detect buildings in urban areas. In the following, the most recent publications in the field of object detection based on combined high-resolution optical and SAR data are summarized. Some of the works presented in this section also contain a three-dimensional part, but their major focus is on two-dimensional building detection. A review of the latest developments concerning building height measurements based on combined optical and SAR data is provided in section 1.3.3.

Fusion of hyper-spectral optical data and one InSAR acquisition

Hepner et al. [1998] and Gamba & Houshmand [2000] propose to jointly use hyper-spectral imagery and InSAR data acquired by airborne sensors to detect and three-dimensionally reconstruct urban areas. After initial co-registration they classify hyper-spectral images into different terrain cover classes and delineate building footprints. Building heights are then assigned by choosing the InSAR value that appears most often within the building footprint. One limitation of these works is low geometric resolution of approximately 20 meters of the hyper-spectral sensor and a horizontal resolution of five meters of the InSAR data. Only very big buildings in urban areas may be detected and height estimation merely works for flat roof buildings.

Combination of one multi-spectral optical image and multi-aspect InSAR data

Xiao et al. [1998] suggest to combine multi-aspect InSAR data with a multi-spectral optical image in order to extract building blocks. They first classify both data separately using a multi-layer perceptron neural network. Then, they combine the two classification results on decision level according to a set of rules in order to suppress false positives. Each pixel is classified into the building or the non-building category. Next, InSAR data of four different aspects are combined to a joint digital surface model (DSM). Based on another set of rules and some morphological operations building regions are extracted and a rectangle is fitted. The results of the first pixel-wise classification of InSAR data and optical image and those of the four combined InSAR aspects are compared in a next step. Those rectangles that contain certain percentage of pixels classified as building by the neural network are decided to be buildings. Finally, building footprints are extracted and the maximum DSM height inside the footprint is interpreted as the building height. This approach contains many parameters to be tuned and is not integrated into a comprehensive probabilistic framework. Building height measurements relying on maximum height inside the building footprint will fail at high buildings that are narrow in range direction. All height information would then be contained in the layover area, which is located outside of the building footprint even if mapped from four different aspects.

Fusion of a multi-spectral optical image and one SAR acquisition

Tupin & Roux [2003] propose an approach to automatically extract footprints of large flat-roofed buildings using one single SAR image and an optical image. The authors first extract double-bounce lines in the SAR image with the ratio line detector proposed in Tupin et al. [1998]. Double-bounce lines occur at the building side that faces the SAR sensor and are part of the building footprint. A projection of the extracted lines to the optical image under the assumption of a known ground height is performed next. Then, edges are segmented in the optical image and filtered. Only those optical edges are kept which are either parallel or orthogonal to the SAR double-bounce line. Rectangles are fitted to the edges based on a set of rules. An alternative for building shapes deviating from rectangles which relies on angular structures (two at each edge) is also presented. This approach relying solely on line features works well in industrial areas characterized by large regularly shaped buildings with flat roofs. It is inappropriate for complex urban scenes containing lots of other object categories in the same scene that occlude or interfere with the buildings.

The method of Tupin & Roux [2003] is extended by Sportouche et al. [2009, 2011]. They combine features found in imagery of high-resolution optical (Quickbird) and SAR (TerraSAR-X) sensors. First, rectangular building footprints are detected in the optical data. Those footprints are refined with additionally extracted edges through a set of rules. Next, the optical building footprints are projected to the SAR image. They are then either validated or rejected based on a classification of the SAR image relying on roof textures, bright lines, and shadows. Building heights are derived simultaneously exploiting the different optical and SAR sensor geometries during a registration of the optical footprints to the SAR image. These works have the same limitations as the ones previously summarized [Tupin & Roux, 2003]. It has only been validated for large flat roof buildings in an industrial area with wide open spaces.

A technique for building recognition in dense urban areas combining line features from mono-aspect InSAR data with classification results from one optical aerial image is presented in Wegner et al. [2009]. Double-bounce lines of buildings are extracted from InSAR data and introduced as features into a classification framework based on a segmentation of the optical image. Optical features and InSAR lines are jointly used in order to evaluate building hypothesis subject to a set of rules. It is shown that the joint use of features derived from optical and InSAR data highly improves the building detection rate and significantly decreases the false positive rate. A slight drawback of this approach is that many parameters have to be tuned manually and anew for a different scene.

Combination of optical imagery and SAR data with a GIS database

Poulain et al. [2008, 2009, 2011] combine high-resolution optical and SAR data with vector data of a GIS database in order to detect changes. No learning step is done, classification exploits prior knowledge and a set of rules. The authors first extract primitives in the images: bright lines in the SAR image and edges, vegetation, shadows, and line segments in the optical image. In the following, they derive features from such primitives and set up a score for each potential building site using Dempster-Shafer evidential theory. Again, this is not a probabilistic approach.

Bottom line

Reconsidering the presented fusion approaches for object detection none actually uses a concise probabilistic framework. All of them are rely on large sets of rules with multiple parameters that have to be tuned manually anew for each dataset. Furthermore, none exploits object-context to support building detection. In the following, state-of-the-art contextual probabilistic approaches are presented.

1.3.2. Classification using context

All approaches presented in the previous section detect buildings merely considering their own appearance. In case other object categories are imaged similarly in the data (e.g., streets, parking lots) or if dealing with highly complex urban scenes it often leads to mistakes. In order to resolve ambiguities between object categories and to improve building detection results, exploitation of contextual knowledge in addition to direct building hints is proposed.

Inclusion of context information into classification of objects in images has its roots in cognitive psychology. Early experimental studies suggest that humans recognize objects based on abstract global information, too, rather than merely on detailed local object information [Potter, 1975;

Palmer, 1975; Biederman et al., 1982]. Palmer [1975] defines the impact of context as the *effects of the environment of an object on the perception of that object, independent of the intrinsic properties of the object itself.* Humans tend to recognize an object not only via its own properties like color, shape, and texture, but as well through its surroundings (i.e., attributes of the entire scene the object may be found in). Oliva & Torralba [2007] provide an overview of the role of context in object recognition and show links between visual cognition, cognitive neuroscience, and computer vision. They state that *contextual influences on object recognition become evident if the local features are insufficient because the object is small, occluded, or camouflaged.*

Various approaches have been proposed in recent years in order to translate findings of cognitive psychology to algorithms and apply them to automated image analysis. A large amount of literature dealing with contextual object detection in imagery exists using non-probabilistic or probabilistic techniques. A non-probabilistic method is, for instance, proposed by Michaelsen & Stilla [2002], who use production nets to group scatterers of industrial buildings in interferometric SAR data[1]. The majority of these methods have their background in computer vision and machine learning. In the following paragraphs focus is on probabilistic contextual methods (Galleguillos & Belongie [2010] provide a comprehensive survey) and particularly on recent developments in Conditional Random Fields for object detection and on few other publications that have influenced this thesis.

Probabilistic contextual approaches

Torralba et al. [2003] propose an approach to categorize terrestrial images into semantic classes completely relying on context information. They extract large scale features of the entire image in order to capture the overall spatial scene structure without processing individual objects or regions. In feature space, spanned by the previously extracted large scale structural features, scenes belonging to the same semantic categories form clusters. Classificaton is then performed with a nearest neighbour classifier. Torralba [2003] extends this method to object detection. He models the relationship between large scale features describing context and object features probabilistically bypassing the identification of context-objects. Murphy et al. [2004] develop the system further for joint object detection and scene classification within a CRF. Both previously mentioned approaches consider context on a global scene level, but do not model relationships between single objects.

Heitz & Koller [2008] exploit implicit context knowledge through what they call "things and stuff" (TAS) approach. The main idea is to, first, cluster image super-pixels based both on local features and their ability to serve as context for objects of interest and, second, to integrate this context prior into a rigorous probabilistic framework for object detection. They combine a window detector for local object detection with context adding predictive power for that particular object category. The TAS idea enables to exploit contextual relations of scenes without having to label all object categories for training beforehand. Benefits of this elegant generic concept are investigated in section 3.1.3 and introduced to our this work.

[1] Probabilites can be integrated into production nets, too, as shown by Michaelsen & Stilla [2002].

Hoiem et al. [2008] propose to make use of contextual knowledge derived from the central viewing perspective of a camera. They probabilistically model the scale and location variance of objects depending on surface orientations and camera viewpoint. The authors show that their approach works well with terrestrial images of urban street scenes. In order to succeed, this method needs a rather simple perspective scene structure, for example, vertical building walls, horizontal flat streets, and sky at the top of the image. It will loose its power if we consider remotely sensed optical images that are usually acquired in nadir perspective leading to only very small perspective changes of different object classes as well as arbitrary object orientation and location.

Classification with Conditional Random Fields

Lafferty and collaborators propose Conditional Random Fields [Lafferty et al., 2001] for labeling sequential data. CRFs are contextual graphical models like Markov Random Fields (MRF), but provide higher modelling flexibilty for classification tasks. Those desirable properties are explained in detail in section 2.2.4. Kumar and Hebert extend CRFs to two-dimensional data and apply them to object detection in images [Kumar & Hebert, 2003, 2006]. They consider contextual knowledge through pair-wise potentials weighted with features. He et al. [2004] learn pairwise relationships between parts of an image at multiple scales. Local, regional and global features are generated and combined within a single CRF. They may thus capture topologies of scenes at various scales from fine details at a very local level to coarse scene structures of the entire image. In Kumar & Hebert [2005] propose a similar approach designing a CRF with two layers. The first layer learns pair-wise relationships between different classes at pixel-level, the second layer captures dependencies between so-called super-pixels[2]. Regions defined by super-pixels are rather large and typically the image is partitioned into approximately twenty super-pixels. This way the CRF can learn both the global distribution of object classes within a scene and local relationships of object class details. This approach works well on small images with clearly observable scene structures consisting of few classes of large objects. In general, CRFs provide a highly flexible framework for contextual classification approaches. Torralba et al. [2005] use Boosting to learn contextual knowledge within a CRF framework. Spatial arrangements of objects in an image are learned by a weak classifier and object detection and image segmentation are done simultaneously. Shotton et al. [2006] propose a similar concept (but relying on features derived from texton maps) they call "TextonBoost" to achieve joint segmentation and object detection applying boosting within a CRF framework.

Learning object class co-occurrences with Random Fields

Another way of directly incorporating contexual knowledge into random fields is to learn whether particular objects or object parts often co-occur in the same scenes and if they have some typical rela-

[2] Three different terms are common in literature to denote irregularly shaped parts of an image that have been aggregated with a segmentation algorithm based on some homogeneity criterion: segments, super-pixels, and regions. The term *region* will be used in this thesis because it is widespread in the remote sensing community.

tion. Characteristic spatial distributions of object classes can directly be captured via co-occurrence matrices as, for example, proposed by Carbonetto et al. [2004]. The authors learn co-occurrences of objects within a Markov Random Field framework. They test their approach on both a regular grid of square image patches and on super-pixels. Rabinovich et al. [2007] propose a similar approach, but formulate a CRF instead of a Markov Random Field. They encode co-occurrence preferences of objects over pair-wise object categories based on image super-pixels. It allows them to distinguish between object categories that often appear together in the same image and, more important, categories that do usually not occur within the same scene. Galleguillos et al. [2010] develop this method further by introducing contextual interactions at pixel-level and at region-level in addition to semantic object interactions via object class co-occurrences. A similar method is proposed by Ladicky et al. [2010] who model object class co-occurrences via an additional potential (that only depends on labels) and add it to the standard CRF energy term. Gould et al. [2008] do not solely rely on occurrences, but add a spatial component by modelling relative locations between two object classes and introducing them into a CRF as a unary potential.

Generalization of node comparisons within a CRF framework

In general, all previously reviewed approaches compare pairs of nodes in the CRF graph structure. Functions relating nodes do not deal with more than two nodes at a time. Kohli et al. [2008, 2009] generalize this classical pair-wise model to higher order potentials that enforce label consistency inside image regions. It allows to model interactions between multiple nodes, functions relate groups of nodes instead of only two. They combine multiple segmentations generated with an unsupervised segmentation method within a CRF for object extraction. Related works of Ladicky et al. [2009] propose a hierarchical CRF integrating features computed in different spatial units as pixels, image regions, and groups of regions. They formulate unary potentials over pixels and regions, pair-wise potentials between pixels and between regions and also a connective potential between pixels and the regions they are contained in. This third potential is an extension of a standard CRF energy function that usually consists of only two parts: The first one contains unary potentials relating labels to data and the second one formulates pair-wise potentials, which compare labels of different nodes including data[3].

Hidden categories in CRFs

Often the variability of object appearances within a single object category is very high. The object category animal, for example, could potentially contain very different kinds of animals from insects to whales making it hardly possible to generate disciminative feature distributions. If we want to circumvent assigning each kind of animal to a separate category, which would then have to be labeled and trained explicitly, hidden categories can be introduced. Quattoni et al. [2007] propose to use

[3] A detailed description of CRFs is provided in section 2.2.4.

hidden object class layers in CRFs and call their method hidden CRF. They assign a vector with a fixed number of hidden subcategories to each object category without training those subcategories explicitly. This method can also prove to be useful in case large objects and their context consist of many small parts (e.g., buildings in urban areas).

CRF remote sensing applications

However, CRFs have only very rarely been used to classify remotely sensed data, yet. Zhong & Wang [2007] set up multiple CRFs to detect settlement areas in an optical satellite image of two meter resolution acquired with the Quickbird satellite. Roscher et al. [2010] use Import Vector Machines (IVM) within a CRF framework to classify regions of two Landsat TM images into multiple land cover classes. They show that their approach outperforms a standard Maximum Likelihood (ML) classifier, a Support Vector Machine, and the IVM without the CRF. Hoberg & Rottensteiner [2010] detect settlement areas in Ikonos images and compare their results to a ML classifier. The CRF facilitates better results because particularly the salt and pepper character of the ML solution is avoided due to the pairwise potentials of the CRF. This approach is extended to the multitemporal case in Hoberg et al. [2010] by adding a third potential to the standard unary and pair-wise potentials of the CRF (cf. [Ladicky et al., 2009, 2010]). In this additional potential the authors model the probability of changes between different land cover types with a transition matrix. Thus, the energy function consists of a term relating labels to data of a node (unary potentials), a term comparing labels of different nodes weighted with data (pair-wise potentials), and a third term which compares labels of the same node in images acquired at different times (transition potential). Settlement areas, for example, are less likely to become vegetated areas, whereas the inverse case is more likely. Lu et al. [2009] use CRFs to extract a digital elevation model from an airborne LiDAR digital surface model. He et al. [2008] apply a CRF to SAR data, with the goal of building extraction, which has been the only time a CRF has been used to classify SAR data so far.

Bottom line

The previously summarized approaches show that contexual probabilistic classification using large databases to automatically learn model parameters, although much used in computer vision and machine learning, has only very rarely been applied to object detection in remote sensing. Only one publication uses CRFs with SAR data, none exists if we look at fusion of SAR and optical data. A major goal of this thesis is to evaluate the potentials benefits and to raise the awareness of the power of CRFs for remote sensing applications.

1.3.3. Building height estimation and reconstruction

Building height estimation and reconstruction often follows building detection. Usually, building footprints are first detected two-dimensionally and the reconstruction step takes it to the third dimension. In order to determine heights and model buildings three-dimensionally from remotely sensed data, various standard methods exist for optical data (e.g., stereo photogrammetry) and SAR sensors (e.g., radargrammetry, interferometric SAR). Photogrammetry, using two or more overlapping high-resolution aerial or satellite images acquired with a certain baseline, is a well elaborated technique developed over the last decades. Applying the stereo-principle to overlapping SAR imagery leads to radargrammetry [Leberl, 1990], which can also be used for building height measurements [Simonetto et al., 2003, 2005; Soergel et al., 2009]. Radargrammetry and optical stereo photogrammetry are not dealt with in this thesis because focus is on the combination of only one optical image with a single SAR acquisition.

Building heights derived from multiple SAR acquisitions

Much research has focussed on the combination of multiple SAR aspects because the success of building height computation is highly dependent on the aspect of the SAR sensor. Hill et al. [2006] and Jahangir et al. [2007] perform building recognition and three-dimensional reconstruction with multiple active-contours evolving simultaneously on radar shadows in multiple SAR images of a scene. A technique for automatic building reconstruction from multi-aspect polarimetric SAR data based on buildings modelled as cuboids within a Maximum Likelihood framework is presented by Xu & Jin [2007]. Besides single SAR images InSAR data of multiple aspects can be exploited [Bolter & Leberl, 2000; Bolter, 2003; Schmitt & Stilla, 2011]. An approach for iterative building detection and reconstruction from multi-aspect InSAR data based on edge and line structures is proposed by Soergel et al. [2003b,a]. Building heights and roof types (flat, gabled, and pent roofs) are estimated by an analysis of the shadow and by fitting planes to the height data. These works have been extended by Thiele et al. [2007a, 2010b] who use InSAR data of two orthogonal aspects for building hypothesis generation. Reconstruction is supported by phase simulations of different building hypothesis and subsequent comparison of the simulated phases to the original InSAR phases. Building heights can also be derived using time series of SAR images with the Persistent Scatterer technique [Ferretti et al., 2000] and multi-baseline approaches [Zhu & Bamler, 2010].

All previously mentioned approaches need more than one SAR acquisition or more than one optical image. Focus in this thesis is on a combination of only one SAR acquisition and one optical image. Nonetheless, all developed methods could be applied to multiple optical or SAR acquisitions, too. One SAR acquisition can either be a single SAR image or an interferometric SAR image pair acquired in single-pass mode. In the following a detailed review of publications that determine heights of buildings only based on one SAR acquisition will be provided. We will need those basic concepts and ideas later on in the approach described in section 3.2. Thereafter, research efforts to combine SAR data and an optical image for building height estimation will be presented.

Building heights via radiometric SAR effects

Franceschetti et al. [2002, 2003] investigate the appearance of buildings in high-resolution SAR imagery by modelling the electromagnetic properties analytically. Depending on the surface roughness either physical optics or geometrical optics are used to model the radar signal return. This method is then extended and applied to building analysis by Guida et al. [2008, 2010]. In real applications such an approach has the principal drawbacks: Geometry, dielectric properties, and roughness of the building have to be known in detail a priori. In other words, access to three-dimensional shape as well as material and surface roughness of the investigated object is required beforehand, which rarely is the case in real world applications. Therefore, focus is on geometrical effects (see section 3.2) and height information contained in radiometric effects is neglected.

A probabilistic parametric model for building reconstruction

Quartulli & Datcu [2004] reconstruct buildings in a single SAR image based on model knowledge. A hierarchical parametric scene model is designed based on prior knowledge. This model is then tested with various parameter settings within a stochastic marked point process. Similar to Markov Random Fields, the objective function consists of two main potentials: a prior potential independent of the data containing a hierarchical parametric scene model and a likelihood potential comparing SAR amplitude distributions on pixel level achieved via training. Results are shown for some few large buildings, but no qualitative accuracy assessment is provided. The authors state that additional data like optical imagery are needed in order to achieve robust results.

Height computation exploiting effects within a single SAR acquisition

Building extraction and height estimation completely relying on radar shadow analysis in a single image is proposed by Bennett & Blacknell [2003]. In general, building height estimation and reconstruction methods based merely on shadow analysis are limited to rural areas or suburban areas. Interfering signal of adjacent objects in urban areas may cause those approaches to fail. Nonetheless, if a shadow is visible, it can provide valuable information, but has to be backed by additional measurements (see 3.2). Cellier and collaborators reconstruct large flat roof buildings from interferometric X-band SAR data of one single aspect [Cellier et al., 2006]. In a first step, they extract features like double-bounce lines and the radar shadows. A mixture model is used in order to resolve different signal contributions from ground, roof, and wall to the phase distribution in the layover area and a first height estimate is achieved. A second height estimate is derived from an analysis of the radar shadow behind the building. The mixture model applied to the interferometric phase data in the layover area is further refined and adapted to full-polarimetric InSAR data in [Cellier & Colin, 2006]. Due to interfering signal of adjacent objects in urban areas, such approach is limited to sparsely distributed large flat roof buildings. Moreover, high-resolution full-polarimetric InSAR data are rarely available.

Combining SAR and optical features for height generation

Only very few research has dealt with combining SAR data with an optical image in order to determine building heights. Only two groups of scientists have focussed on this topic, yet. Tupin [2003] determines the heights of flat-roofed industrial buildings analysing the layover area in a single SAR intensity image. First, a building map is generated manually from an optical aerial image which defines expectation areas for line detection at particular buildings. Bright lines are extracted in the SAR intensity image and heights are computed exploiting three-dimensional information contained in layover (cf. section 3.2.1) with a set of rules. It relies on a very simple building model and geometrical considerations of radar viewing geometry. Tupin & Roux [2005] regularize a height model derived by means of radargrammetry within regions of an aerial photo. First, the optical image is segmented into homogeneous regions. They second generate a region adjacency graph and a Markov Random Field is set up based on the graph. A specially designed potential function replaces the usually used Potts model in the prior term. The assumption is made that heights within a homogeneous region of the optical image tend to be similar. Additionally, heights of different image regions also should be similar in case no strong gradient in the optical image separates both regions. On the contrary, a height jump should occur between two regions if they are separated by a high gradient. Reconsidering that they use radargrammetric heights the authors actually combine two SAR acquisitions with one optical image. A similar approach dealing with the same configuration as this thesis is developed by Denis et al. [2009]. They extend the method of Tupin & Roux [2005] to three-dimensionally reconstruct an urban area from high-resolution InSAR data and an optical image. In addition, they propose a graph-cuts-based inference method for energy minimization of the MRF. They perform tests of separate and joint likelihood functions of amplitude and phase data. Optical data is introduced via the prior term of the MRF, where the gradient magnitude serves as an indicator for height discontinuities (similar to [Tupin & Roux, 2005]).

Simulation-based height determination combining SAR and optical data

Brunner et al. [2008, 2010] propose an iterative simulation and matching approach to compute single building heights. They manually generate a simplified three-dimensional CAD-model for each building by visual analysis of optical remote sensing images. Those CAD-models are fed into a SAR image simulator which generates SAR reflectivity maps for varying building heights (all shape parameters stay fixed). All resulting simulated images are compared to the original SAR images and a similarity score is calculated via Mutual Information (in order to take into account different grey value statistics of the simulated reflectivity maps and original SAR images). The height parameter leading to the best match is considered to be the closest to the true building height. This method delivers good results for single isolated buildings, but the authors state it can hardly cope with closely located buildings and other objects leading to interfering radar signal returns.

Bottom line

Most of the state-of-the-art approaches described combine different means of building height determination for reasons of robustness. Additionally, they rely on features being extracted in a preprocessing step in order to introduce some high level model knowledge we have about the object of interest and the radar sensor viewing geometry. Different height estimates are combined based on sets of rules that often assume very simple building shapes.

Two main disadvantages arise from those approaches: First, the proposed methods are often particularly designed for a specific scene and cannot be transferred to a different one without major changes. Second, the accuracy of a single building height is usually not given. Thus, a highly flexible least squares adjustment framework is proposed combining different height estimates computed on feature level and assigning an accuracy value to each building height.

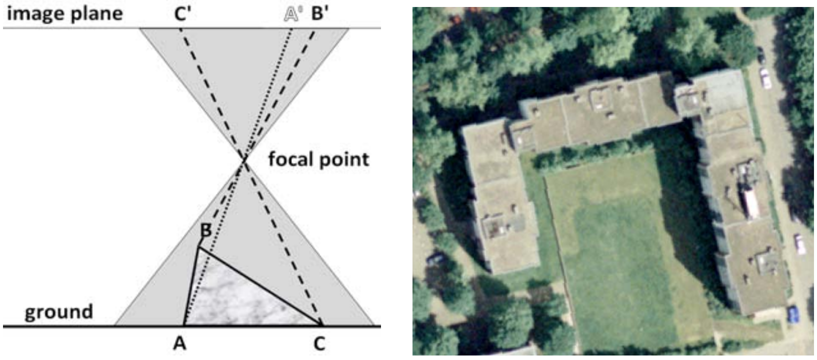
2. Basics

In this chapter, the basic theory of the used data (2.1) and of applied probabilistic methods (2.2) are explained. In the first section, the focus is on the differences between optical images and SAR data in terms of geometry and radiometry. General properties of optical images are discussed before turning to SAR data and the technique of SAR interferometry (InSAR). Only the basic concepts and principles will be provided as a reminder. Then, it is described how buildings are mapped by optical and SAR sensors. Furthermore, those effects are highlighted that are strong hints at buildings and can thus serve as features in the classification framework.

In the second section, methodological foundations of the probabilistic classification framework applied here are laid. The reader is first reminded of fundamentals and rules of probability theory. Second, the general principles of graphs and how they can be used to represent context are introduced. The link between graphs and probabilistic approaches is explained and the general concept of graphical models is derived. Then, Markov Random Fields are discussed, the current state-of-the-art probabilistic contextual method that has been widely applied to a large variety of pattern recognition tasks. Next, a detailed description of *Conditional Random Fields* highlighting conceptual differences to MRFs is given. Details of training and inference procedures are given in Appendix A.

2.1. Sensors

SAR and optical sensors make use of very different measuring techniques. An object that is imaged by both sensors will appear quite dissimilar in terms of geometry and radiometry. At first glance this complicates the analysis of an object based on joint use of SAR and optical data. However, this sensor configuration also provides complimentary views of the same object. In case one sensor is not capable of acquiring features that discriminate the object category of interest from others it may well be possible by mapping the same object with an entirely different technique. Optical images, for example, provide a rich source of information in terms of color and texture. But if we cannot distinguish buildings with grey untextured roofs from adjacent streets and parking lots, another way of viewing the same object might help. This is where the SAR technique is benefitial since it gathers particular features of buildings that do not occur at streets or parking lots. In the following, principle differences between SAR and optical acquisition systems will be described with emphasis on the complimentary ways buildings appear in both data.



(a) (b)

Figure 2.1.: Mapping of ground object by an optical sensor with a central perspective: (a) Schematic view, (b) distorted building in an aerial image

2.1.1. Optical sensors

Only main properties of optical remote sensing data will be pointed out because this topic has already extensively been dealt with in literature. For details the reader is referred to Campbell [2002], Kraus [2007], and Lillesand et al. [2008], for example.

Optical sensors are passiv devices. They receive electromagnetic signal of nanometer wavelength emitted by the sun and reflected towards the sensor by objects on ground. That is why they cannot take images at night or if clouds are covering the area of interest. They are capable of acquiring multi-spectral object information in the visible and infrared spectra. Today's spaceborne high-resolution optical sensors provide resolution below half a meter. Airborne sensors achieve even higher resolution down to several centimeters. Two or more images of the same scene can be used to automatically create detailed three-dimensional models through matching techniques and bundle block adjustment. Object positions are measured via directional measurements similar to human perception. The part of the electromagnetic spectrum of sunlight being reflected by an object highly depends on the object's material. An optical sensor primarily captures chemical properties of an object on the ground (whereas SAR sensors capture physical object properties like the conductivity). Since reflected sunlight is received, shadows of objects are mapped, too. They are often good indicators of elevated objects like buildings or trees. If in cities shadows are too long they can hamper the interpretability of the images. In addition, haze can blur images and thus a compromise between small shadows (at noon) and less impact of haze (early in the morning) has to be found. Many optical satellites pass over Germany between ten and eleven in the morning.

All reflected sunlight received by the sensor propagates through a system of optical lenses and through the focal point. Figure 2.1 shows how an object on the ground is mapped onto the image plane. The schematic view (Fig. 2.1(a)) assumes a central perspective valid for standard frame cameras as opposed to line scanners, which have a central perspective only orthogonally to their flight direction. Central perspective leads to distortions of objects in the image, this means growing object height and increasing horizontal distance to the nadir point of the sensor on the ground lead to more distortions. Building facades directed towards the sensor are visible while the roof top partly falls over to the opposite side (cf. Fig. 2.1(b)). This effect can also be seen in figure 2.1(b). Point A is not mapped in the image because it is occluded by B. Later on these distortions will be exploited in order to measure building heights (3.2).

2.1.2. SAR sensors

Only a short summary of the basic geometric and radiometric properties of SAR imaging systems is provided because exhaustive literature already exists and the reader is referred to Leberl [1990], Meier et al. [1993], Raggam et al. [1993], and Soergel [2010] for a comprehensive review.

In contrast to optical cameras radar sensors are active devices. They emit a pulse signal in the microwave domain which is then reflected at some object and received by the sensor. Due to longer wavelength and active sensor principle mentioned, radar sensors are capable of mapping objects at night and through cloud coverage (depending on wavelength). These properties make them a suitable tool for a wide range of applications, for example in the military domain or for rapid mapping of destructions after the occurrence of natural disasters. Furthermore, compared to optical sensors radar signals cover a far narrower bandwidth of the electromagnetic spectrum and the wavelength is much longer. Instead of measuring directions they measure the distance between sensor and object. As a result of this so-called slant range measurements and long wavelength, radar sensors are sensitive to physical properties like roughness or conductivity of an object. The signal power received by the sensor depends on sensor design, distance, and backscattering properties of the object like geometric shape, directivity, and reflectivity. Resolution in range direction δ_{sr} is a function of pulse length τ and the velocity of electromagnetic waves c (Eq. 2.1) divided by two because the signal travels from sensor to object and back. We have to consider the sensor's viewing angle θ to get the ground range resolution δ_{gr}.

$$\delta_{sr} = \frac{c\tau}{2} \quad , \quad \delta_{gr} = \frac{c\tau}{2\sin\theta} \tag{2.1}$$

The azimuth resolution of real aperture radar (RAR) is diffraction limited and usually defined as $\delta_{ra} = \theta_a R \approx \frac{\lambda R}{d}$. It depends on the sensor's beamwidth in azimuth θ_a and on the distance R to the ground. The beamwidth can be approximated with the antenna length d in azimuth (i.e., the real aperture) and the signal wavelength λ. This is a principle drawback of RAR because the distance between sensor and ground is very long, the wavelength is in microwave domain and using extremely large apertures (i.e., very long antennas) is not feasible. State-of-the-art imaging radar sensors thus use a *synthetic aperture*. Synthetic aperture radar (SAR) sensors synthetically combine many low-resolution (in azimuth direction) RAR acquisitions in flight direction that greatly overlap. An object on the ground is illuminated multiple times by the sensor as long as it is contained in the radar footprint (in azimuth direction). Those multiple low-resolution measurements are combined to generate one high-resolution image by integrating all echos of an object of all RAR acquisitions in

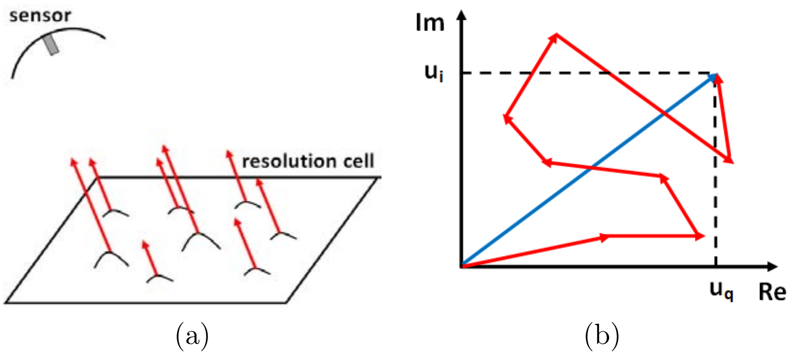


Figure 2.2.: SAR signal sum within one resolution cell: (a) reflected signal at multiple scatterers, (b) complex cartesian representation of the SAR signal (single contributions in red, sum signal in bue)

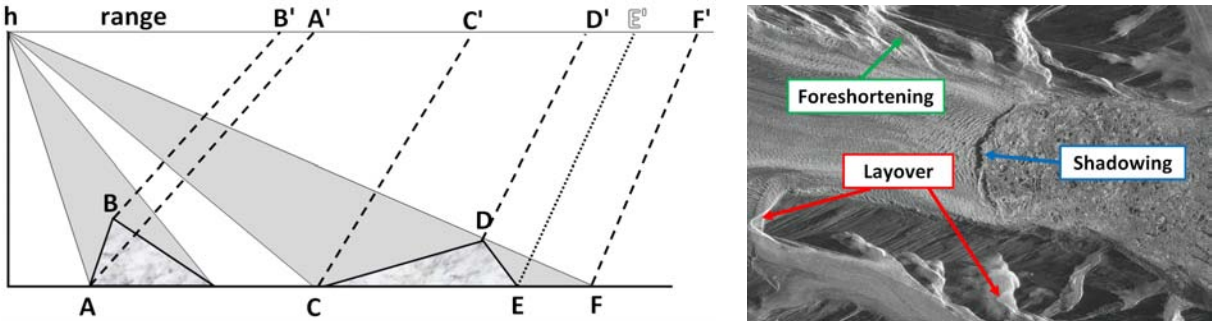
azimuth. This greatly improves the azimuth resolution δ_{SARaz}, which is then completely independent of the sensor's distance to an object and its wavelength. It can be approximated with half of the synthesized antenna length $\delta_{SARaz} = \frac{d}{2}$. δ_{SARaz} improves with decreasing antenna length d because a shorter antenna leads to a larger footprint on the ground, which means that an object is illuminated more often (i.e., the synthetic aperture gets longer).

The SAR signal u is complex-valued $u = u_i + ju_q$ with a real part u_i and an imaginary part u_q (cartesian coordinates in Fig. 2.2). The final pixel value of a SAR image is the sum (blue arrow in Fig. 2.2) of multiple coherent signal reflections N (red arrows in Fig. 2.2) on the ground (Eq. 2.2, a_n: amplitude, ϕ_n: phase). The standard model, being valid for most scenes of natural land cover, assumes the presence of many independent scatterers within one resolution cell contributing to the final signal received by the sensor. The fact that the pixel value is the coherent sum of a large number of complex signals also leads to the speckle effect, which causes a grainy appearance of regions of homogeneous land cover. Even though speckle is no noise, but the signal its effects may be considered as nuisance of the underlying "pure" backscatter. In this sense speckle is often modelled to act as source of multiplicative disturbance. This contradicts the common model for optical images, where additive noise occurs. Therefore, it is inappropriate to apply edge or line detectors developed for optical images directly to SAR data. One way to deal with speckle is to work with detectors that are based on ratios of grey values. The most frequently used state-of-the-art line detector relying on ratios and providing a constant false alarm rate is proposed by Tupin et al. [1998]. Adaption of this line detector to double-bounce line extraction will be described in section 2.1.4.

$$u_i = Re\{u\} = \frac{1}{N}\sum_{n=1}^{N} a_n \cos\phi_n \ , \quad u_q = Im\{u\} = \frac{1}{N}\sum_{n=1}^{N} a_n \sin\phi_n \tag{2.2}$$

The slant range perspective leads to several effects carrying valuable information about the three-dimensional shape of an object. Three different effects occur: shadowing, foreshortening, and layover (Fig. 2.3(a)). An example image of a mountaineous area in Antarctica acquired with the TerraSAR-

X satellite in Stripmap mode is shown in Fig. 2.3(b). Although A is located in front of B, it is mapped behind B' because of the shorter distance between the sensor and B. Foreshortening occurs between points C and D because the plane between them is tilted towards the sensor and thus C' and D' are mapped closer together. Point D occludes E leading to a shadow area between D' and F' without E' being mapped. Exploiting those effects to determine building heights will be explained in 3.2.



(a) (b)

Figure 2.3.: Geometric SAR effects: (a) Foreshortening, layover, and shadowing, (b) TerraSAR-X Stripmap image of 3 m resolution (range direction from left to right) of the Larsen ice shelf in Antarctica (©DLR)

2.1.3. InSAR

Interferometric SAR (InSAR) has been widely described in literature and only a brief overview of the basic concepts shall be provided here. A comprehensive review can be found in [Bamler & Hartl, 1998; Hanssen, 2001].

Several processing steps have to be conducted in order to generate a height model from two SAR images. Both SAR images are taken from slightly different acquisition positions and viewing angles. The SAR image pair has to be co-registered in two steps: first coarsely with a precision of several pixels followed by fine adjustment with an accuracy of one tenth of a pixel. Then, the images are oversampled in order to avoid aliasing and phase differences are computed by pixel-wise complex multiplication. As a result we get phase differences in a range between 0 and 2π.

The maximum height Δh that fits into the 2π range is a function of the signal wavelength λ, the distance R between sensor and object, the viewing angle θ, and the orthogonal baseline $B_\perp$ between the two sensor positions (p is one for single-pass and two for dual-pass interferometry).

$$\Delta h \approx \frac{\lambda \cdot R \cdot \sin(\theta)}{p \cdot B_\perp} \tag{2.3}$$

As a rule of thumb, a smaller orthogonal baseline leads to a larger height range being contained within the 2π range, but also to less accurate height measurements (Eq. 2.4). Usually, height variations of the terrain (e.g., in mountaineous areas or cities) are greater than the unambiguous height of the 2π range. Therefore, we often have to apply a so-called *phase-unwrapping* step in order to resolve the phase to height ambiguity. Many approaches exist and phase-unwrapping may be done if we consider only smooth and continuous terrain variations. However, if turning to terrain with large, abrupt height changes and frequent discontinuities, like in urban areas, those phase-unwrapping is hard to resolve.

Additionally, the height accuracy σ_h decreases with an increasing range R between sensor and object, with longer wavelengths λ, and with worse signal-to-noise ratios SNR (L: number of looks). In case of a single-pass configuration decorrelation caused by temporal and atmospheric effects can be neglected. Thermal noise occurs and is included in the equation of height accuracy σ_h via the SNR.

$$\sigma_h \approx \frac{\lambda \cdot R \cdot \sin(\theta)}{p \cdot 2\pi \cdot B_\perp \cdot \sqrt{SNR} \cdot \sqrt{L}}. \tag{2.4}$$

2.1.4. Mapping of buildings in optical images and SAR data

In this section focus is on how buildings are mapped by different sensor types. Emphasis is on basic hints[1] for buildings in urban areas. From those hints features are derived that will be used within a CRF framework for building detection in Chapters 3 and 4. In addition, some of the hints will be exploited for building height estimation combining optical and SAR data in section 3.2. In the following, it will be explained how buildings appear in optical imagery before turning our attention to SAR data.

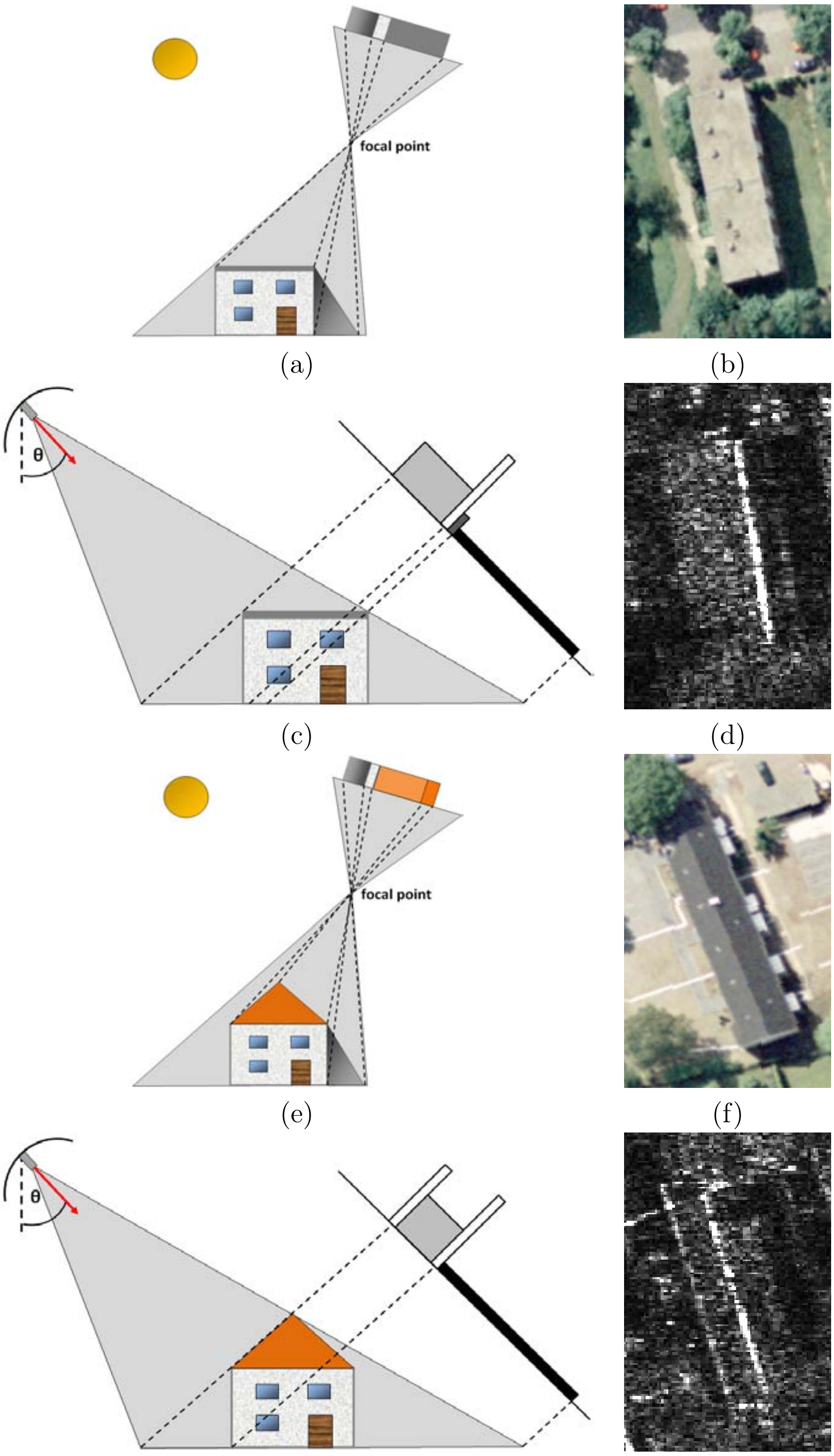
Optical sensors can acquire multi-spectral object information and measure directions as opposed to SAR sensors measuring ranges. Buildings are mapped completely different in optical data (cf. Fig. 2.4). Geometry of mapped building parts in an optical image depends on the central perspective model, the perspective projection according to camera orientation. Parts of facades are visible if a building is not located directly in the nadir of a sensor like shown in figures 2.4(a,b,e,f). The roof will be slightly shifted away from the nadir point. This perspective effect carries valuable information about the height of a building as well as the shadow of the sun. Due to the central perspective, the gable roof side that is tilted away from the sensor (left roof side in Fig. 2.4(e,f)) is shortened in the image. A roof plane tilted away from the sun is mapped slightly darker in the image (right roof part in Fig. 2.4(e,f)). In terms of color information building roofs are usually greyish (Fig. 2.4(b,f)), reddish, or brownish. They mostly appear as rather homogenous areas in the image that may comprise also small super-structures like chimneys and dormers (cf. Fig. 2.4(b,f)).

[1] Object hints in data and features used as input to feature vectors in a classification framework are distinguished in this thesis. Hints characterize and indicate a particular object category in data. For example, the double-bounce line in SAR data and the sun shadow in the optical image are hints at buildings. Features are scalar values derived from object hints. Multiple features can be generated based on one hint like mean, maximum, and median intensity values within a shadow area of a building.

The way buildings appear in SAR data has already been described in much detail in literature. First investigations of single-, double-, and triple-bounce effects of radar signal at buildings were published by Dong et al. [1997] and related publications. Thiele et al. [2007b, 2010a,b] comprehensively discuss the appearance of different building types. Works presented in this section are based on their findings and the extraction algorithm for double-bounce lines has been developed in collaboration with Antje Thiele [Wegner et al., 2009].

Building mapping in SAR data depends on the oblique illumination of the scene, that is the image projection in slant range geometry. Furthermore, it depends on sensor parameters, on properties of the imaged object itself, and on the local environment of the object. Figures 2.4(c,d,g,h) show how flat roof and gable roof buildings are mapped by a radar sensor schematically and as magnitude SAR image. The SAR magnitude profile of a building typically is a sequence of several types of signal returns: layover, double-bounce line between ground and building wall, roof, and radar shadow. Layover (cf. 2.1.2) is caused by single-bounce reflection of signal at the building and ground (or other objects) in front of it. In figures 2.4(c,g) the layover area is illustrated in light grey in the magnitude profile. This corresponds to the inhomogeneous grey area left of the white line in the real SAR image in figure 2.4(d) and the area between the two bright lines in figure 2.4(h), respectively. Objects in front of the building with respect to the range direction cause signal return that interferes with the signal retrieved directly from the building. In case of the flat roof building trees, a footpath, and a grass area can be seen in the optical image (Fig. 2.4(b)). Signal return of the building originates from direct reflection of the roof and the facade. A layover area is situated closest to the sensor in the image of the flat roof building because its range is the shortest. It ends at the bright double-bounce line, which is caused by the reflections at a dihedral corner reflector spanned by ground and wall along the building. All signal is mapped to one single line because wall and ground enclose an angle of 90 degrees and thus all distances have equal length. It has a high magnitude (cf. marked as red dot in magnitude profile in Fig. 2.5(a)) because all double-bounce reflections of the entire building facade and the ground in front are collected. This line coincides with a part of the building footprint and can be distinguished from other lines of bright scattering using the InSAR phases as we will see later on in this section. An area of weak signal return caused by direct reflection of parts of the roof occurs behind the double-bounce line, which is shown in dark grey with low magnitude in figure 2.4(c). Dealing with gable roof buildings (Fig 2.4(g,h)) a second bright line parallel to the double-bounce line, located in front of it, can be present. All single-bounce signal of the near-range roof side is collected in one line because the roof plane normal and incoming radar signal span an angle of almost zero degrees, that is all ranges have approximately the same length. The characteristics of this signal contribution and of the ones previously described heavily depend on the sensor's aspect, its viewing angle, reflectance properties of the mapped building, and its local vicinity (cf. [Thiele et al., 2010b] for details). Ground behind the building is partly occluded leading to a dark region in the SAR image due to missing signal return at the corresponding ranges.

A building also leads to specific patterns in interferometric phase data because the phase value of a single range cell results from a mixture of backscatter of different contributors like ground, facade,



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 2.4.: Mapping of a flat roof building and a gable roof building: (a,e) schematic optical sensor view, (b,f) corresponding optical images, (c,g) schematic SAR sensor view, (d,h) corresponding SAR magnitude images (range direction left to right)

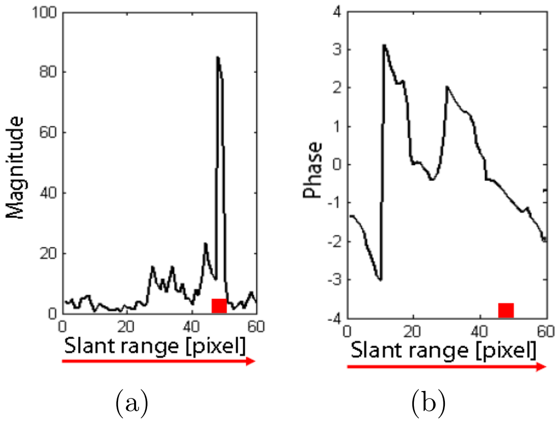


Figure 2.5.: Profile of magnitude and InSAR phase data at the position of the double-bounce line (red dot) [Wegner et al., 2009]

and roof in the layover area. A phase profile of a flat roof building is shown in figure 2.5(b). Again, the appearance is characterized by a layover region left of the double-bounce line. A homogeneous roof region right of the double-bounce line is absent in figure 2.5(b) due to the narrow building width. The phase value at the position of the double-bounce line has a similar phase value as the ground (phase at ground height is zero in this case). In the shadow area behind the building no signal is received and the related InSAR phase carries no useful signal, but noise only.

Layover and shadow areas are often hard to extract automatically in urban areas due to interfering signal of adjacent objects. The most prominent building hint in SAR data is the double-bounce line, thus the focus is on it for building detection. However, it should be noted that its intensity depends on the aspect of the sensor, investigated together with additional radiometric aspects by, for example, Brunner et al. [2009] and Guida et al. [2010]. In terms of geometrical precision and accuracy it is shown in Wegner et al. [2010] that double-bounce lines can be extracted with an accuracy of about one pixel. In the following the double-bounce line extraction approach, jointly developed with Antje Thiele [Wegner et al., 2009], is explained.

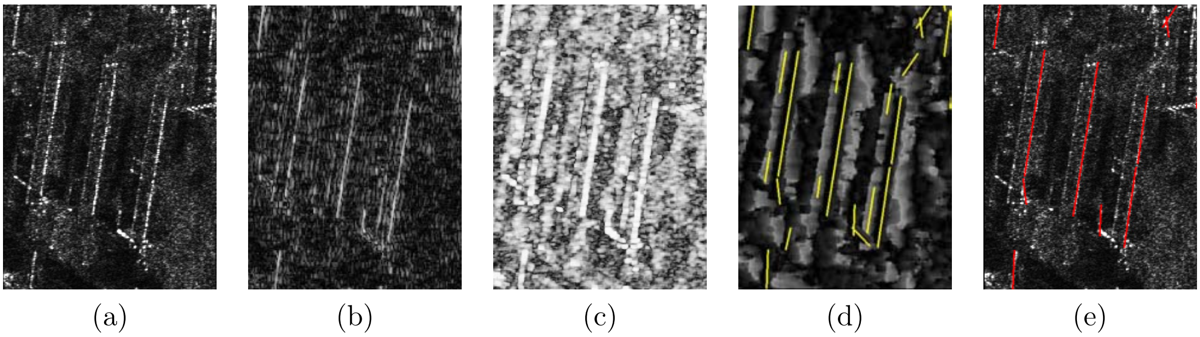


Figure 2.6.: Double-bounce line extraction in InSAR data (range from left to right): (a) amplitude image, (b) line probabilities, (c) coherence, (d) height differences, (e) filtered lines overlaid to amplitude image (©Intermap Technologies) [Wegner et al., 2009]

First, bright lines are segmented in the magnitude data. Based on this set of lines, only the ones caused by a dihedral corner reflector spanned by ground and building wall are used as building hints. In order to exclude all lines that do not fulfil this criterion, the local InSAR heights are analysed. Finally, the filtered double-bounce lines are projected into the same ground range geometry as the optical data. As previously discussed, bright double-bounce lines are very useful hints at buildings because they provide information about the true location of a part of the building footprint. The full process of double-bounce line extraction is shown in Fig. 2.6. Line extraction is carried out in slant range geometry based on one of the original amplitude images (Fig. 2.6(a)) using an adapted ratio line detector according to Tupin et al. [1998]. This template detector determines the probability of a pixel belonging to a line. Here, eight different template orientations are considered. The probability image of the vertical template orientation is shown in Fig. 2.6(b). Short line segments are fitted to straight lines and edges, respectively, by linear approximation and subsequent prolongation (yellow lines in Fig. 2.6(d)).

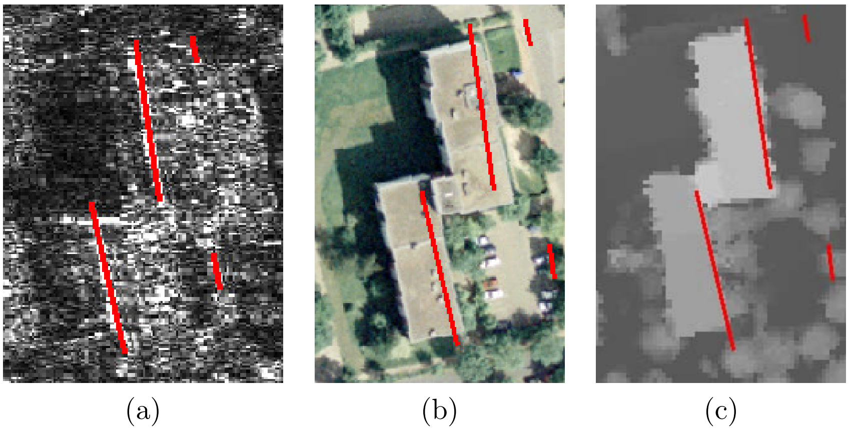

(a) (b) (c)

Figure 2.7.: Extracted double-bounce lines superimposed on (a) an amplitude image of the InSAR image pair (range from right to left), (b) an orthophoto, (c) and a digital surface model derived from airborne laserscanning data

After line extraction, the interferometric heights are computed. Local InSAR heights are investigated at each line in order to discriminate lines caused by direct reflection and lines due to double-bounce reflection between either ground and wall or roof and substructures. For this filter step, the height difference between digital surface model (DSM) and digital terrain model (DTM) is used. The DSM corresponds to the calculated InSAR heights. A filtering step is done in order to generate a DTM. Only DSM pixels with a high coherence value and an InSAR height close to the global mean terrain height are considered. A DTM height value is then calculated over an area of 50 m x 50 m in ground range geometry. Then, height differences (i.e., normalized DSM) between DSM and DTM are calculated (Fig. 2.6(d)). In the following line filtering step, lines are considered to be double-bounce lines of buildings if their neighbouring pixels show a low mean height difference value. The filtered double-bounce lines are displayed as red lines overlaid to the amplitude image in Fig. 2.6(e). Finally, these double-bounce lines are projected to the optical image using the InSAR heights. The resulting position of the double-bounce lines of a flat roof building (Fig.

2.7(a)) superimposed onto the optical image is displayed in figure 2.7(b). We can observe that the roof of the building in the optical image overlaps with the double-bounce line due to the perspective distortion induced by the optical sensor (cf. Fig. 2.4(a,b)). This effect can be exploited to measure the building height combining the SAR double-bounce line and the roof overlap in the optical image as will be explained in section 3.2.1. A corresponding DSM derived from airborne laserscanning data overlaid with the lines (Fig. 2.7(c)) shows that they have a high horizontal positioning accuracy. We can rely on double-bounce lines both for building detection and for building height estimation.

2.2. Probabilistic modelling

Now the fundamentals of probabilistic modelling for classification tasks will be introduced. Step by step it will be explained how graphical models can be used to express probabilities of data and labels. These considerations will lead us to the basics of context formulation within Markov Random Fields. Then, we will have learned all necessary theory to turn our attention to Conditional Random Fields and corresponding learning and inference techniques. In this section and the following sections and chapters data are denoted with x and labels with y. Labels are the categories that we want to automatically assign to the data, for example, building or non-building in this thesis[2].

2.2.1. Fundamentals of Probability Theory

In this section a short reminder of the basic concepts of probability theory is given. A very comprehensive explanation can be found in Bishop [2006, chap. 1]. The two basic rules of probability theory are the *sum rule* (Eq. 2.5)[3] and the *product rule* (Eq. 2.6). All further approaches and ideas presented in the following sections are based on these two fundamental rules.

$$P\left(x\right) = \sum_{\mathbf{y}} P\left(y, x\right) \tag{2.5}$$

$$P\left(y, x\right) = P\left(x|y\right) P\left(y\right) \tag{2.6}$$

The sum rule states that if we have a joint probability $P\left(x, y\right)$, we can compute the marginal probability $P\left(x\right)$ by summing all possible configurations of $\mathbf{y}$ of the joint probability $P\left(y, x\right)$ (Eq. 2.5). Often, it is not convenient to directly model the joint probability $P\left(y, x\right)$, but express it as a product of the conditional probability $P\left(x|y\right)$ and the marginal probability $P\left(y\right)$ instead. Considering the symmetry property $P\left(x, y\right) = P\left(y, x\right)$ we derive the *Bayes Theorem* (Eq. 2.7), one essential concept of pattern recognition and computer vision. Reconsidering the product rule

[2] Variables representing scalar values are written in normal face type, vectors as lower-case letters of bold face type, and sets or matrices in capital letters of bold face type

[3] It should be noted that if we write $\sum_{\mathbf{y}} P\left(y, x\right)$ we have to, first, evaluate the joint probability $P\left(y, x\right)$ for all possible configurations of labels $\mathbf{y}$ and, second, to sum all evaluated joint probabilities.

(Eq. 2.6) we can write $P(y|x)\,P(x) = P(x|y)\,P(y)$. Division through $P(x)$ leads to the Bayes Theorem. It allows us to express posterior probabilities $P(y|x)$ in terms of conditional probabilities $P(x|y)$ and marginal probabilities $P(y)$ and $P(x)$. This will turn out to be highly convient when constructing posterior probabilities with graphical models for large inference tasks.

$$P(y|x) = \frac{P(x|y)\,P(y)}{P(x)} \tag{2.7}$$

The conditional probability $P(x|y)$ is often called the likelihood, whereas the marginal probability $P(y)$ of labels is a prior. For example, setting $P(y)$ to a uniform distribution (i.e., all labels are equally likely to occur) and maximizing the likelihood would lead to a standard Maximum Likelihood classifier. Marginal probability $P(x)$ is expressed making use of the sum rule and the product rule. We substitute the joint probability $P(y,x)$ in equation 2.5 with the right side of equation 2.6. Probability $P(x)$ can then be written as the sum of all products of conditional probability $P(x|y)$ and marginal probability $P(y)$ considering all possible states of y:

$$P(x) = \sum_{\mathbf{y}} P(x|y)\,P(y)\,. \tag{2.8}$$

Marginal probability $P(x)$ is often expressed as a function $Z(x)$, which is called the *partition function*. Basically it is a normalization constant for a given data set and turns the product $P(x|y)P(y)$ into a probability with values between zero and one. Later on we will need it for Markov Random Fields and Conditional Random Fields.

2.2.2. Graphical Models

In this section the idea of probabilistic graphical models is introduced. It will be explained how image data can be structured with graphs and the link to probabilistic modelling is described. For details the reader is referred to Bishop [2006, chap. 8], for example. In order to represent structures in data, we have to find some way to represent relationships between small parts of it, which partition the data (i.e., divide it gapless into non-overlapping areas). Considering images, a small part may be a pixel, a square patch, or an irregular region generated by an image segmentation algorithm (Fig. 2.8). Instead of simply determining properties of all image parts separately and then classifying each, regardless of its adjacent parts, prior knowledge contained within each parts vicinity shall be introduced and learned. In other words, each small part of an image should have an influence on its local neighbours and maybe even on all other parts. The inverse case should also hold namely that an image part is possibly influenced by all other parts of the image. An approach fullfilling these requirements will enable us to learn and infer context at any scale, from very local structures to global scenes. A powerful concept allowing us to represent relationships between small parts of images and meet the aforementioned requirements is a *graph*.

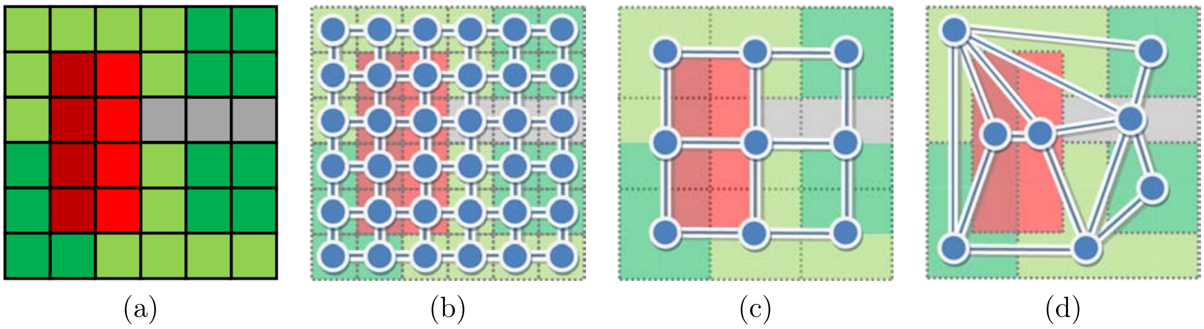
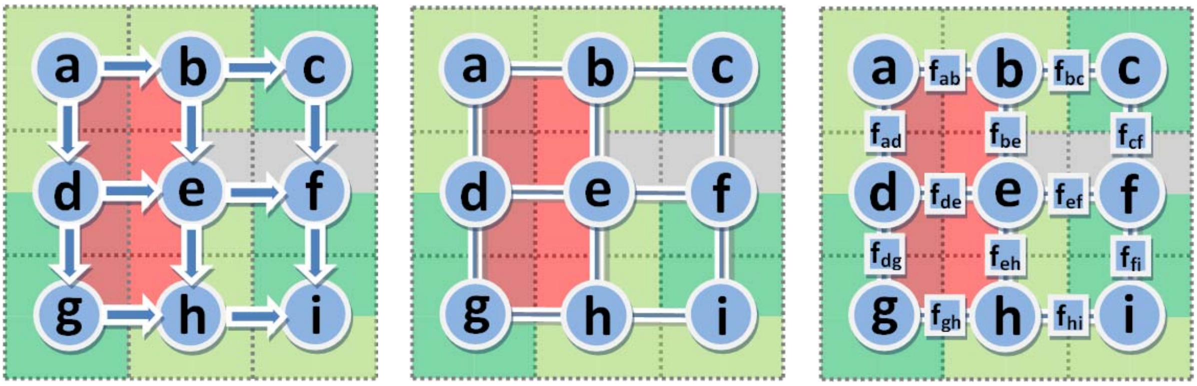


Figure 2.8.: An image represented with three different graph structures: (a) artificial scene containing one gable roof building (light and dark red), low vegetation (light green), high vegetation (dark green), and a driveway (grey), (b) graph on pixel-level, (c) graph on patch-level, and (d) graph on region-level; dotted lines represent the spatial extent of each node, nodes are shown as blue circles with white boundaries, and edges linking nodes are drawn as blue lines with white frames

Graphs are an approach to represent a set of entities (image parts in our case) somehow related. In general, the entire set of entities is called a graph $\mathbf{G}$, a single entity is called a node n, and the link between two entities is called an edge e. The set of all nodes n is $\mathbf{N}$ and $\mathbf{E}$ is the set of all edges e. Vice versa each node n is a member of the set $\mathbf{N}$ of all nodes ($n \in \mathbf{E}$) and each edge e is a member of the set of all edges $\mathbf{E}$ ($e \in \mathbf{E}$). $\mathbf{G}$ consists of $\mathbf{N}$ and $\mathbf{E}$, which can be expressed as $\mathbf{G} = \mathbf{N} \cup \mathbf{E}$ or $\mathbf{G} = (\mathbf{N}, \mathbf{E})$. The spatial unit represented via a node is the smallest entity to be labeled and the aim is to decide whether a node is labeled as building or non-building. A spatial unit of an image could be a single pixel, a square image patch, or an image region. Examples of those three possibilities and the corresponding graphs are shown in figure 2.8. Spatial units represented by a single node are enclosed in dotted lines, nodes are represented as blue circles with white boundaries linked with edges of the same colors. In figure 2.8(a) an artifical scene containing a gable roof building (light and dark red), low vegetation (light green), trees (dark green), and a driveway is shown. Pixels are separated with black lines, which means we have an image of $6 \times 6 = 36$ pixels. A first possibility is to assign a node to each pixel (Fig. 2.8(b)) resulting in a very high number of nodes and edges. In case of big images training (cf. section A.1) and inference (cf. section A.2) become unfeasible and thus single pixels are usually aggregated to larger square patches regardless of the scene content (Fig. 2.8(c)). Another possibility to reduce the number of nodes and edges, but still respecting object boundaries, is to assign a node to each image region generated by segmentation (Fig. 3.1.1(d)). This third way of representing an image with a graph has certain advantages which will be described in detail in section 3.1.1. It will now be explained how a graph representation can be exploited for probabilistic modelling.

Graphical models are probabilistic models based on the general concept of graphs. They represent mathematical operations like multiplication and sums of conditional and marginal probabilities (see Eq. 2.5 and 2.6) via graphs. The joint probability $P(x, y)$ of an entire dataset is formulated in terms of products of conditional and marginal probabilities of adjacent nodes within a graph. This



(a) (b) (c)

Figure 2.9.: Different graph types: (a) directed graph, (b) undirected graph, (c) and factor graph; dotted lines represent the spatial extent of each node, nodes are shown as blue circles with white boundaries, edges as blue lines with white frames.

decomposition of a joint probability distribution into products of factors of locally adjacent nodes is called *factorization*. The way a joint distribution *factorizes* describes the way it is represented through local conditional and marginal probabilities. Two different kinds of graphical models are usually distinguished: directed (Fig. 2.9(a)) and undirected graphical models (Fig. 2.9(b)). Directed graphical models are often referred to as *Bayesian Networks* and undirected graphical models are also called *Random Fields*. Both can be represented in terms of a factor graph [Kschischang et al., 2001] (Fig. 2.9(c)), a convenient way to express edges via functions f that depend on the nodes they link. All three example graph types in figure 2.9 are based on the patch graph structure of figure 2.8(c) for means of simplicity. All following considerations are generally valid for any kind of graph structures (e.g., the ones of figures 2.8(c,d)).

In order to explain the idea behind using graphs to represent probability distributions, Bayesian Networks are used as an example (Fig. 2.9(a))[4]. A Bayesian Network is a directed graphical model as shown in Fig. 2.9(a). The joint distribution $P(a, b, c, d, e, f, g, h, i)$ of all nine nodes of the graph can be expressed with the product of their corresponding marginal and conditional probabilities through the product rule (Eq. 2.6). We can thus write the joint distribution of the directed graph shown in figure 2.9(a) as:

$$\begin{aligned} P(a, b, c, d, e, f, g, h, i) = {} & P(a)\, P(b|a)\, P(c|b)\, P(d|a)\, P(e|b, d) \\ & P(f|c, e)\, P(g|d)\, P(h|e, g)\, P(i|f, h)\,. \end{aligned} \tag{2.9}$$

We see that the probability of node a does not depend on other nodes leading to the marginal probability $P(a)$. It is called a parent node of b and d because both nodes depend on a, whereas b and d are child nodes of a. The probabilities of all nodes besides a depend on other nodes (represented

[4] A detailed explanation of undirected graphical models is provided in the next section 2.2.3.

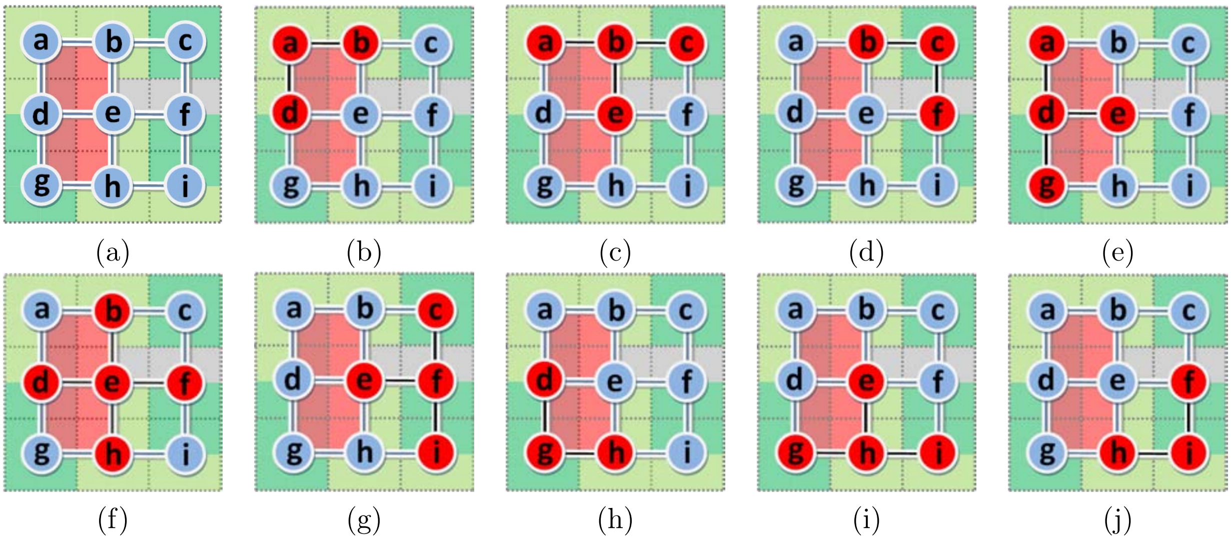
with arrows in the graph (Fig. 2.9)), we have conditional probabilities. All nodes except a and i are child nodes and at the same time parent nodes because each of them depends on other nodes, but also is a parent node of at least one node. Since no other node depends on node i, it is a child node of f and h, but no parent node. We can exploit these directed relationships and the corresponding conditional probabilities to express some kind of prior knowledge we have about the dependencies between nodes. Any joint distribution of a directed graphical model can be factorized into this product of conditional probabilities, which explicitly represent the dependencies between single adajacent nodes in the graph. Considering this hierarchical structure of parent nodes and child nodes, we can generally state that using the product rule repetitively any Bayesian Network with K nodes factorizes into a product of factors $P\left(k|\pi\left(k\right)\right)$ where $\pi\left(k\right)$ is the parent node of node k (Eq. 2.10). This is the essential link between graphs and probabilistic models. It will be exploited for incorporation of context in section 3.1.

$$P\left(\mathbf{x},\mathbf{y}\right)=\prod_{k=1}^{K}P\left(k|\pi\left(k\right)\right) \tag{2.10}$$

We have to consider the conditional independence property if dealing with graphical models. This concept will become essential turning to Markov Random Fields (2.2.3) and Conditional Random Fields (2.2.4). In a directed graphical model as shown in figure 2.9(a), the two nodes b and d are conditionally independent. Both depend on a, but not on each other, which becomes obvious if we decompose the joint distribution of b and d into a product of conditional probabilities $P\left(b,d|a\right)=P\left(b|a\right)P\left(d|a\right)$ using the product rule (Eq. 2.6). Both nodes statistically depend on a, but they are statistically mutually independent. Thus, a prerequisite of directed graphical models is that they must not contain any cycles. In case the criterion is violated, the conditional independence assumption does not hold any more. If following the arrows in the graph, we must not be able to visit a particular node more than once. Reconsidering the example mentioned previously and shown in Fig. 2.9(a), a cycle would be present if we would add an arrow pointing from node e back to node a. This would lead to node a being visited three times: as a parent node, via b and e, and via d and e. If we need to formulate dependencies leading to cycles within the graph structure we have to turn to undirected graphical models as shown in figure 2.9(b) and explained in the following section 2.2.3.

2.2.3. Markov Random Fields

Markov Random Fields (MRF) are graphical models set up on undirected graph structures. They are generative models estimating the joint distribution $P\left(\mathbf{x},\mathbf{y}\right)$ of data $\mathbf{x}$ and labels $\mathbf{y}$, which can be decomposed into a product of factors $P\left(\mathbf{x}|\mathbf{y}\right)P\left(\mathbf{y}\right)$ (cf. Eq. 2.6). The main difference to Naive Bayes is that MRFs explicitly model the prior, marginal probability $P\left(\mathbf{y}\right)$ of labels $\mathbf{y}$, via the Markov property.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 2.10.: Decomposition of the undirected graph (a) into conditionally independent neighbourhood sets of nodes in a 4-neighbourhood system with cliques of second order (b-j) (nodes in one neighbourhood set in red)

In order to derive a probability distribution of an undirected graph (Fig. 2.10(a)), we have to reconsider the conditional independence assumption introduced in the previous section. In a 4-neighbourhood system as shown in figure 2.10(a) all nodes except those at the image boundaries have four neighbours (here only node e has four neighbours). The Markov property states that each node only depends on its neighbours. Every node is only conditioned on its direct neighbours and it is conditionally independent of all other nodes in the graph. Figures 2.10(b-j) show how the undirected graph of figure 2.10(a) is decomposed into neighbourhood sets of nodes (red nodes linked with black edges). Node a, for instance, conditionally depends on b and d (Fig. 2.10(b)). A link between nodes a and i would violate the MRF conditional independence assumption.

In order to derive the joint probability distribution of MRFs, the concept of cliques has to be introduced. Cliques $\mathbf{c}$ are defined as a subset of adjacent nodes of a graph that are linked pair-wise with an edge. For example, pair-wise cliques of the neighbourhood of node e are $\mathbf{c}(e,b)$, $\mathbf{c}(e,d)$, $\mathbf{c}(e,f)$ and $\mathbf{c}(e,h)$ (Fig. 2.10(f)). Cliques of a graph can be used in order to derive a probability distribution via the Hammersley-Clifford theorem of Markov Random Fields [Besag, 1974; Clifford, 1990]. This theorem links MRFs to Gibbs distributions, allowing the derivation of a global probability distribution from the sum of local clique potentials of neighbouring nodes[5]. It is valid for any kind of graph structure.

The distribution $P(\mathbf{y})$ of all labels $\mathbf{y}$ is represented by the product of *potential functions* $\psi_{\mathbf{C}}(\mathbf{y}_{\mathbf{c}})$ over the pair-wise cliques $\mathbf{C}$ of the graph (Eq. 2.11). Normalization is needed to transform potentials to probabilities, which is done by division through the *partition function* Z. It does not have to be conducted after each node potential update, but after each training iteration.

[5] A detailed proof of the Markov-Gibbs equivalence established by the Hammersley-Clifford theorem is out of the scope of this thesis. It is given, for example, in [Li, 2009, p. 28]

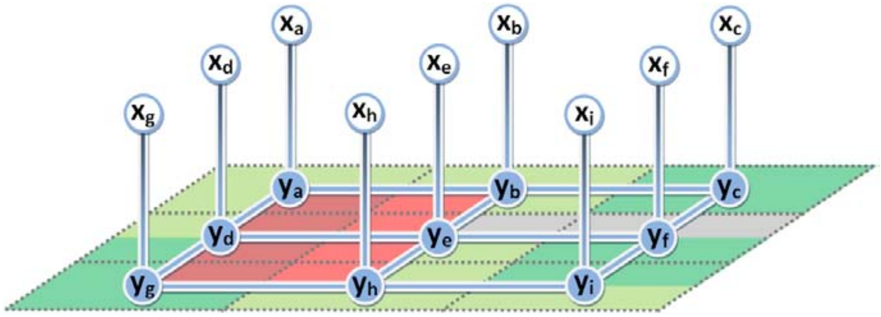


Figure 2.11.: MRF labeling of nodes with labels $\mathbf{y}$ that depend on data $\mathbf{x}$

$$P(\mathbf{y}) = \frac{1}{Z(\mathbf{y})} \prod_{\mathbf{C}} \psi_{\mathbf{C}}(\mathbf{y_c}) \ , \quad Z(\mathbf{y}) = \sum_{\mathbf{y}} \prod_{\mathbf{C}} \psi_{\mathbf{C}}(\mathbf{y_c}) \tag{2.11}$$

In contrast to Bayesian Networks potential functions do not necessarily have to be probabilities. In Bayesian Networks each factor at a node is a marginal probability or a conditional probability conditioned on its parent node. Random Fields are not restricted to this choice of potential functions. They are formulated in a more flexible way allowing for the design of arbitrary potential functions at a node [Bishop, 2006, chap. 8.3]. Usually, the potential functions $\psi_{\mathbf{C}}(\mathbf{y_c})$ are expressed with functions out of the exponential family $\psi_{\mathbf{C}}(\mathbf{y_c}) = \exp(E(\mathbf{y_c}))$ because exponential functions have the advantage of never being zero or negative for any clique. We can design specific energy functions $E(\mathbf{y_c})$ in order to model some prior knowledge for image classification.

A MRF for image classification has two different kinds of cliques. In figure 2.11 label y_a is linked two labels y_b and y_d, for example. Another type of edge exists between the data x_a of node a and its label y_a. Therefore, the standard energy term for image classification with a MRF framework can be written as the sum of two parts: the first one relates labels and data, whereas the second formulates dependencies between labels of neighbouring nodes (Eq. 2.12). To ease notation the node of interest is called i and the one in its neighbourhood $\mathbf{N}_i$ it is compared to is j. $\mathbf{S}$ is the set of all nodes in the graph. The neighbourhood of a particular node is defined as shown in figure 2.10.

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{i \in S} \log P_i(x_i|y_i) + \sum_{i \in S} \sum_{j \in N_i} \beta y_i y_j \right) \tag{2.12}$$

In the Bayesian context the first term in Eq. 2.12 can be viewed as the likelihood, whereas the second one is a prior over labels $\mathbf{y}$. Likelihood $P_i(\mathbf{x}_i|\mathbf{y}_i)$ uses data only from a single node i, not from all other nodes (like CRFs). The prior (the right term in Eq. 2.12) only compares adjacent labels y_j to the investigated label y_i, usually using (for binary classification) the Ising model $\beta y_i y_j$ (with penalty value β). Partition function $Z(\mathbf{x})$, which can be interpreted as the distribution $P(\mathbf{x})$ of data $\mathbf{x}$ in the Bayesian framework, acts as a normalization constant (for a given data set). It can be expressed as the sum of all possible label configurations of the product $P(\mathbf{x}|\mathbf{y})\,P(\mathbf{y})$.

MRFs have been widely applied to image interpretation in general and to object detection in remotely sensed data in particular. In Klonowski & Koch [1997] and related works, for example, an MRF is set up on image regions for object detection in urban areas based on optical aerial images, whereas Tupin & Roux [2005] use MRFs to regularize a radargrammetric height model.

Nonetheless, the standard MRF framework has some limitations: In order to satisfy the conditional independence assumption, a label y_i of a node can only be compared to the label y_j of its direct neighbour. In addition, it is impossible to consider data if comparing labels of two nodes because an edge exists only between data x_i of a particular node and its corresponding label y_i (cf. Fig. 2.11). This implies we cannot consider data of other nodes to decide for label y_i. Conditional Random Fields are no subject to these restrictions (although closely related to MRFs) as we will see in the next section and thus provide a higher flexibility in terms of energy function formulation.

2.2.4. Conditional Random Fields

In this section the concept of Conditional Random Fields, which belong to the family of undirected graphical models, will be introduced. They are closely related to Markov Random Fields, but differ in some important properties, which will be explained here. CRFs were originally introduced by Lafferty et al. [2001] to label one-dimensional text sequences. Kumar & Hebert [2003, 2006] extended CRFs to two-dimensional data to label images[6]. Lafferty et al. [2001] define CRFs as ($\mathbf{x}$ contains all observations and $\mathbf{x}$ all labels):

> *Let $\boldsymbol{G} = (\boldsymbol{N}, \boldsymbol{E})$ be a graph such that $\boldsymbol{y} = (\boldsymbol{y}_n)_{n \in \boldsymbol{N}}$, so that $\boldsymbol{y}$ is indexed by the vertices of $\boldsymbol{G}$. Then $(\boldsymbol{x}, \boldsymbol{y})$ is a conditional random field in case, when conditioned on $\boldsymbol{x}$, the random variables $\boldsymbol{y}_n$ obey the Markov property with respect to the graph: $P(\boldsymbol{y}_n | \boldsymbol{x}, \boldsymbol{y}_w, w \neq n) = P(\boldsymbol{y}_n | \boldsymbol{x}, \boldsymbol{y}_w, w \sim n)$, where $w \sim n$ means that w and n are neighbours in $\boldsymbol{G}$.*

Thus, a CRF is an undirected graphical model (i.e., random field) globally conditioned on all observations $\mathbf{x}$. It has several desirable properties to make it a highly flexible and efficient framework for contextual probabilistic classification:

- It is a discriminative model,
- the conditional independence assumption is relaxed,
- global observations may be incorporated in the association potential,
- the interaction potential is a function of both: labels and observations,
- labels and observations in the interaction potential are not limited to the local neighbourhood, but may be regarded globally.

[6] Kumar & Hebert [2003] call their method Discriminative Random Fields because they use discriminative functions for both the unary and the pair-wise potentials. This particular choice of the potential functions does not change the general CRF framework and thus we will keep the notation Conditional Random Field in this thesis.

We will now see what these porperties actually mean and what the differences are compared to standard MRFs as described in the previous section. CRFs are based on the maximum entropy approach, which is known to be able to provide accurate and robust classification results [Nigam et al., 1999][7]. They are discriminative techniques meaning that they directly model the posterior distribution $P(\mathbf{y}|\mathbf{x})$ of the labels $\mathbf{y}$ given data $\mathbf{x}$ as a Gibbs distribution, which leads to a relaxation of the conditional independence assumption with respect to MRFs. This results in a much higher flexibility in terms of context formulation compared to MRFs, which are generative methods to model the joint probability $P(\mathbf{x}, \mathbf{y})$ via a Gibbs distribution (cf. 2.2.3). It implies that features can be computed in spatially overlapping units in contrast to MRFs.

An example to exploit this property would be to segment an image in multiple scales and write features of all scales to a node representing a region at the highest scale. A neighbouring node at highest scale would also receive the same features of coarser scales because it is contained in the regions of smaller scales, too. We cannot do this in a standard MRF framework if we strictly follow the theory because it would violate the conditional independence assumption (cf. 2.2.3). Nodes in the MRF graph would not be statistically independent any more due to overlapping regions. In addition, CRFs are globally conditioned on all data, thus we can design potential functions relating data of arbitrary locations in an image. In figure 2.12 a CRF of a subset of the example graph is shown. For instance, label y_d of node d is not only connected to its own data x_d, but also to the data of all other nodes x_a, x_b, x_d, and x_e (cf. 2.11). We cannot do this with MRFs because a node would then be a function of arbitrary nodes in the graph and not only of its direct neighbours leading to a violation of the conditional independence assumption. Moreover, we can also compare labels of arbitrary nodes in the graph as opposed to MRFs where only node labels of adjacent nodes can be compared.

Another important property of CRFs results from the previous considerations: In the prior term we can compare node labels with respect to data, too. We are no longer limited to pure label comparisons (cf. prior term of MRF in Eq. 2.12), but we can incorporate data. If we can learn dependencies of data $\mathbf{x}$ and of labels $\mathbf{y}$ globally without any restrictions on node locations in the graph, we gain a highly flexible tool for contextual modelling[8]. In the following, it is described how these properties can be expressed more formally.

We have an energy term $E(\mathbf{x}, \mathbf{y})$ encapsulating unary and pair-wise parts (cf. 2.2.3). Potential functions of CRFs do not necessarily have to be formulated as probabilities, but they have to be valued positively. Usually, functions out of the exponential family are used to turn the energies into potentials. In order to gain a posterior distribution $P(\mathbf{y}|\mathbf{x})$, we need to turn potentials into probabilities by normalizing them through the partition function $Z(\mathbf{x})$. We may then write the posterior distribution $P(\mathbf{y}|\mathbf{x})$ as:

[7] The idea of maximum entropy is to prefer the most uniform distribution satisfying all constraints learned from training data. Considering discriminative classification the posterior $P(\mathbf{y}|\mathbf{x})$ is the distribution to be learned. A more detailed explanation of links between probability theory and maximum entropy principle is given by Guiasu & Shenitzer [1985], for example.

[8] This flexibility is exploited to propose new ways of contextual learning in section 3.1

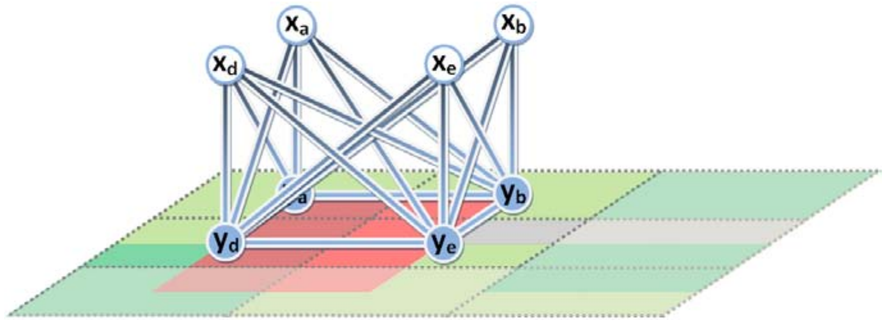


Figure 2.12.: CRF labeling of nodes with labels $\mathbf{y}$ that depend on all data $\mathbf{x}$ globally (only a subset of the nodes is shown for visualization purposes)

$$P\left(\mathbf{y}|\mathbf{x}\right) = \frac{1}{Z\left(\mathbf{x}\right)} \exp\left(E\left(\mathbf{x},\mathbf{y}\right)\right). \tag{2.13}$$

Following the notations of Kumar & Hebert [2006] we can express the energy term $E\left(\mathbf{x},\mathbf{y}\right)$ as the sum of a first term that associates labels with data $A_i\left(\mathbf{x},y_i\right)$ and a second term that defines how labels interact (incorporating data) $I_{ij}\left(\mathbf{x},y_i,y_j\right)$:

$$E\left(\mathbf{x},\mathbf{y}\right) = \sum_{i\in\mathbf{S}} A_i\left(\mathbf{x},y_i\right) + \sum_{i\in\mathbf{S}}\sum_{j\in\mathbf{N}_i} I_{ij}\left(\mathbf{x},y_i,y_j\right) \tag{2.14}$$

Substituting this energy function into equation 2.13 we get the standard CRF expression for two-dimensional data of the posterior $P\left(\mathbf{y}|\mathbf{x}\right)$ of labels $\mathbf{y}$ conditioned on all data $\mathbf{x}$ [Kumar & Hebert, 2006]:

$$P\left(\mathbf{y}|\mathbf{x}\right) = \frac{1}{Z\left(\mathbf{x}\right)} \exp\left(\sum_{i\in\mathbf{S}} A_i\left(\mathbf{x},y_i\right) + \sum_{i\in\mathbf{S}}\sum_{j\in\mathbf{N}_i} I_{ij}\left(\mathbf{x},y_i,y_j\right)\right). \tag{2.15}$$

The left term of equation 2.15 is also called *association potential* $A_i\left(\mathbf{x},y_i\right)$. It measures how likely a node i is labeled with y_i given all data $\mathbf{x}$. $I_{ij}\left(\mathbf{x},y_i,y_j\right)$ is also referred to as the *interaction potential* and it defines how the labels of two nodes i and j interact. As previously explained, both potentials have access to the whole image. In particular the interaction potential $I_{ij}\left(\mathbf{x},y_i,y_j\right)$ is not only a function of adjacent labels y_i and y_j in the local neighbourhood (like in case of MRFs, compare Eq. 2.2.3), but of all data $\mathbf{x}$, too. Neighbourhood $\mathbf{N}_i$ of node i may potentially be the entire image. This is convenient if we want to compare labels based on underlying data. In addition, both the association potential and the interaction potential are defined over all data, the entire orthophoto and all SAR data in this work. Hence, we can introduce both local and global context knowledge, which is a major advantage concerning automatic analysis of high-resolution remote sensing data of urban areas. To obtain a posterior probability $P\left(\mathbf{y}|\mathbf{x}\right)$ of labels $\mathbf{y}$ conditioned on data $\mathbf{x}$, the exponential of the sum of association potential and interaction potential is normalized by division through the partition function $Z(\mathbf{x})$ (Eq. 2.16). It has to be evaluated for each new parameter set during training, but is a constant for a given data set once parameters have been adjusted.

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} \exp\left(\sum_{i \in \mathbf{S}} A_i(\mathbf{x}, y_i) + \sum_{i \in \mathbf{S}} \sum_{j \in \mathbf{N}_i} I_{ij}(\mathbf{x}, y_i, y_j)\right) \tag{2.16}$$

Association potential $A_i(\mathbf{x}, y_i)$ and interaction potential $I_{ij}(\mathbf{x}, y_i, y_j)$ can be formulated in various ways. For example, Hoberg & Rottensteiner [2010] use a Maximum Likelihood classifier for $A_i(\mathbf{x}, y_i)$ to generatively determine the most likely label of node i. Nonetheless, the choice is not limited to probabilistic methods, thus Support Vector Machines could be inserted, too, for instance. This high degree of flexibility also applies to the comparisons of labels via the interaction potential $I_{ij}(\mathbf{x}, y_i, y_j)$. Carbonetto et al. [2004]; Rabinovich et al. [2007], for example, learn co-occurrences of different object categories globally over an entire database of images. Our modelling of both $A_i(\mathbf{x}, y_i)$ and $I_{ij}(\mathbf{x}, y_i, y_j)$ is closely related to the approach proposed by Kumar & Hebert [2006]. Both potentials are discriminatively formulated as linear models:

$$A_i(\mathbf{x}, y_i) = y_i \mathbf{w}^T \mathbf{h}_i(\mathbf{x}), \tag{2.17}$$

$$I_{ij}(\mathbf{x}, y_i, y_j) = y_i y_j \mathbf{v}^T \boldsymbol{\mu}_{ij}(\mathbf{x}). \tag{2.18}$$

Vector $\mathbf{h}_i(\mathbf{x})$ contains all node features (e.g., those derived from the segmented SAR double-bounce lines of section 2.1.4). Those features are scalar values, mean and maximum intensity within an image region that is represented with a single node in the graph, for example. Vector $\mathbf{w}^T$ contains the weights of the features in $\mathbf{h}_i(\mathbf{x})$ that are tuned during the training process (A.1). Features that help to discriminate the object classes receive high weights, whereas those that do not considerably contribute are down-weighted. In the interaction potential $\mathbf{I}_{ij}(\mathbf{x}, y_i, y_j)$ the comparison of labels y_i and y_j follows the Ising model $\beta y_i y_j$ (because we deal with a binary classification task with the two categories building and non-building). With $\beta = 1$, the product $y_i y_j$ becomes -1 if labels y_i and y_j do not belong to the same class, whereas their product is 1 in case both labels are equal. As already stated by Korč & Förstner [2008], vector $\boldsymbol{\mu}_{ij}(\mathbf{x})$ enables to support or suppress this term and ignoring $\boldsymbol{\mu}_{ij}(\mathbf{x})$ would lead to the standard MRF smoothing potential, the traditional Ising model (cf. prior in Eq. 2.12). In our case, the edge feature vector $\boldsymbol{\mu}_{ij}(\mathbf{x})$ is simply calculated by subtracting the feature vectors of nodes j from such of the node i of interest $\boldsymbol{\mu}_{ij}(\mathbf{x}) = |\mathbf{h}_i(\mathbf{x}) - \mathbf{h}_j(\mathbf{x})|$. However, in general $\boldsymbol{\mu}_{ij}(\mathbf{x})$ could also be chosen based on other features than such already used for the association potential and other methods of comparing the features (e.g., concatenating the feature vectors of two nodes) are possible, too. Best results are achieved with absolute differences of node features. Vector $\mathbf{v}^T$ contains the weights of the edge features being adjusted during training [9]. In general, it is convenient to view the standard CRF as a factor graph (cf. Fig. 2.9(c)) where feature functions relating nodes are contained in the interaction potential (Eq. 2.18).

[9] CRF parameters are automatically learned via training with semantically annotated data and inference is needed for computation of node labeling probabilities (cf. Appendix A).

3. Methodology

In this chapter a detailed explanation of new ideas and methods is given. The focus is twofold: Novel expressive formulations of contextual knowledge with CRFs are introduced (3.1) and concepts to exploit the different mapping properties of high-resolution optical and SAR sensors for building height estimation are proposed (3.2).

First, basic considerations regarding contextual information in remote sensing data and differences to typical computer vision scenes are discussed. The need for more sophisticated contextual learning is motivated. CRFs on image regions generated by segmentation are explained, changes compared to square image patches concerning graph structure and message passing during inference are addressed. Thereafter, two novel approaches are given, which enable a more sophisticated context integration into CRFs: one via the pair-wise term (3.1.2), another one via the unary term (3.1.3).

The second part is dedicated to the accuracy of height estimation based on one SAR acquisition and an optical image (section 3.2). Building height measurements are introduced exploiting the different mapping properties of both sensor types (3.2.1). Separate and merged measurements are discussed. Multiple heights per building are combined in a least squares adjustment framework, weighted according to their accuracies, and a final estimated height is assigned to each building. Testing results of all methods are shown in Chapter 4.

3.1. Formulation of context with Conditional Random Fields

Most approaches making use of context information have been developed to analyze terrestrial images. Lots of them are inspired by research in cognitive psychology and neuroscience dealing with human perception of the environment. A great part of the success of integrating context into object detection is owed to relatively simple, but expressive information. Considering figure 3.1(a), for example, sky usually is in the upper half of an image, whereas streets are in the lower half; buildings always stand upright, roofs are located above facades. Such simple assumptions are not valid anymore or have to be relaxed if dealing with remotely sensed data. Regarding optical remote sensing data, the perspective changes compared to terrestrial images. Humans who are not familiar with such data often encounter difficulties interpreting scenes acquired in nadir view (Fig. 3.1(c)) because they do not correspond to their usual perception of objects. Scene interpretation gets complicated if not only the perspective changes, but also the sensoring technique itself as in case of SAR (Fig. 3.1(d)). Unlike optical sensors, which measure directions in the visible domain of the

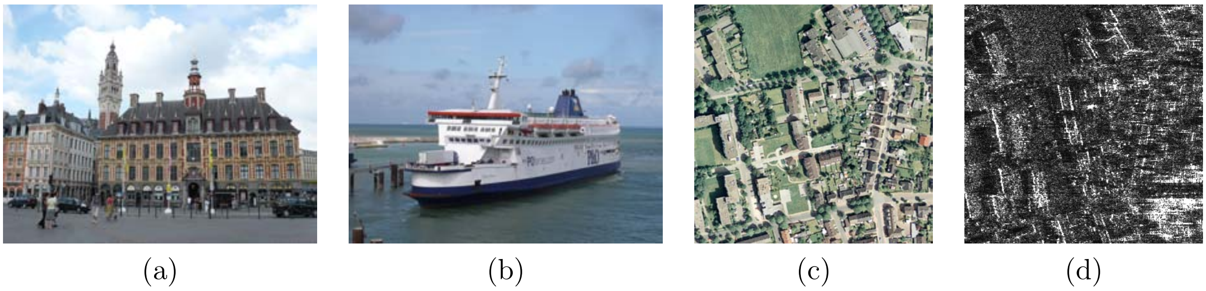

(a) (b) (c) (d)

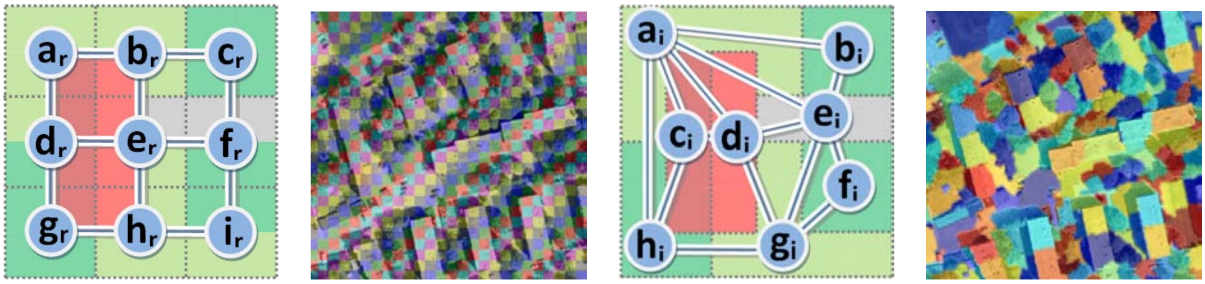
Figure 3.1.: Comparison of typical images for object detection (a, b) in computer vision (building facade, ship), (c) in optical remote sensing, and (d) radar remote sensing (buildings).

electro-magnetic spectrum, SAR sensors measure slant-ranges in the microwave domain. Dissimilar features characterize objects, three-dimensional effects as layover lead to displacements. These radiometric and geometric mapping differences prevent direct pixel-wise fusion of high-resolution SAR and optical images. Fusion based on features, extracted in the original geometries and projected to a common coordinate system, is therefore proposed in this thesis. This concept implies that novel assumptions about context modelling and its contribution to object detection have to be made.

One way to really fully exploit global context within a CRF framework is proposed by Rabinovich et al. [2007]. They set up a co-occurence matrix of multiple object categories learning whether certain objects often occur in the same image. Galleguillos et al. [2010] extend this method by adding contextual interactions at pixel-level and region-level to the CRF energy term similar to Ladicky et al. [2010]. Learning co-occurrences is well adapted to object detection in usual computer vision images where only few object categories appear together in an image. Water and ships jointly occur (Fig. 3.1(b)), but water and cats do not, for example. Often, a single instance of each category covering a large area is present per image (e.g., ship and building facade in Fig. 3.1(a,b)).

Several aspects are different if dealing with remote sensing data. A high number of object categories exists in urban scenes and multiple instances per category occur covering small areas in the images. An urban scene of the city Dorsten, Germany, mapped by an optical aerial camera is shown in figure 3.1(c). Buildings are distributed over the entire image, each consisting of a relatively small number of pixels. No simple ordering like "above" and "below", but complex patterns of object categories exist. For example, buildings are usually aligned with streets, driveways connect street with building, and shadow directly neighbours each building (always on the same side in one acquisition).

We have seen in section 2.2.4 that the standard interaction potential basically is a smoothing term. Its degree of smoothing is steered by the edge features. It is well adapted to large single contiguous objects like building facade and ship (Fig. 3.1(a,b)), but tends to over-smooth if many small objects with narrow gaps in-between are contained in a scene (Fig. 3.1(c)). Separating gaps between buildings are potentially misclassified. This effect increases in cases where building and adjacent non-building regions are similar in terms of features, which calls for novel ways of context formulation besides standard case and co-occurrences.



(a) (b) (c) (d)

Figure 3.2.: (a) Regular graph on image patches in a 4-connectivity neighbourhood, (b) image patch grid overlaid to aerial photo, (c) irregular graph on image segments, (d) image segmentation (with multi-scale Normalized Cuts [Cour et al., 2005]) overlaid to the same optical aerial image as in (b)

3.1.1. Conditional Random Fields on image regions

MRF and CRF are usually based on graphs with regular structures motivated by the two-dimensional grid of digital images (cf. Fig. 2.8(a)). Representing each pixel of an image with a node in the graph is infeasible for large images and datasets because training and inference become computationally very expensive. A standard principle to reduce graph size and computational costs in computer vision is to divide an image into a grid of square image patches (e.g., [Kumar & Hebert, 2006; Vishwanathan et al., 2006a; Zhong & Wang, 2007]). Each patch aggregates several pixels and is represented with nodes as shown in figure 3.2(a). Four pixels are aggregated to a patch, which is represented with a node in the graph. This graph still has a regular grid architecture, nodes are linked with edges to four adjacent nodes (node e_r in Fig. 3.2(a)) in a 4-neighbourhood system. Exceptions are nodes in image corners (a_r, c_r, g_r, and i_r) being linked to two neighbouring nodes and those at image borders (b_r, d_r, f_r, and h_r) with three edges each. Although reducing computational costs such square patches have disadvantages particularly with respect to remote sensing data. A patch grid is set up independently of the scene content following the image grid structure. It does not consider objects contained in the image, therefore patches often cut across boundaries as can be seen in figure 3.2(a). The image patch of node e_r contains four different object categories: Street (light grey), grassland (light green), and building roof (light red). A patch grid overlaid to a part of a real optical aerial image is shown in figure 3.2(b). Patches do not preserve object boundaries, but contain pixels of buildings and other objects. It is the finest representation of the data and each patch will be labeled either building or non-building in its entirety. Reconsidering the patch of node e_r in figure 3.2(a) this leads to half of the pixels being misclassified in any case. In addition, feature distributions characterizing both classes lose their discriminative power due to a single node possibly representing mixtures of classes. As a consequence, expressiveness of the classification suffers because a decision surface in feature space separating classes cannot be adjusted appropriately during training. Although patches reduce the number of nodes compared to pixels, graphs are still big. Large homogeneous regions are partitioned into many patches, too, because the regular graph grid does not consider the underlying scene structure.

Principle drawbacks of the patch graph can be overcome by introducing a graph based on image regions. It preserves object boundaries, the structure of the scene is expressed via the graph structure, its size is usually significantly reduced thus decreasing computation time [Wegner et al., 2011c], and expressive context formulation is facilitated. Graphs on image regions have already been used with MRFs (e.g., [Klonowski & Koch, 1997; Tupin & Roux, 2005]). He et al. [2004] were the first to combine graphs of image regions with CRFs for object detection. More sophisticated contextual learning based on image regions was published by, for example, Kohli et al. [2008, 2009] and Gould et al. [2008, 2009]. All region-based approaches proposed so far have been applied to computer vision data, none exists for object detection in remote sensing data.

The ideal case of a graph based on regions, which perfectly match object boundaries, is shown schematically in figure 3.2(c). Pixels of same colour are aggregated to regions separated by dotted lines. A single node is assigned to each region that now represents an object or object part. Nodes do no longer contain mixtures of objects. Region-graphs call for a particular treatment because they have an irregular structure, defined by the segmentation (Fig. 3.2c,d), where nodes have different numbers of neighbours as opposed to the regular patch grid. Node a_i is linked to five other nodes (b_i, c_i, d_i, e_i, and h_i), whereas b_i is only linked to two nodes (b_i and e_i). It should be noted that an additional advantage of image regions is that they capture object shapes enabling the introduction of features like shape, size, main orientation, and roundness that can potentially serve as features to discriminate buildings from their environment (Fig. 3.2(d)).

Introduction of an irregular graph implies changes regarding inference. Within a graph, messages between nodes are passed along edges (blue lines with white boundaries in Fig. 3.2(a,c)). A node both sends and receives messages from an adjacent node via the same edge. In case of a regular grid of image patches all nodes have equal numbers of neighbouring nodes (except those at image boundaries and in corners) and they exchange the same number of messages via edges. Depending on the connectivity of the neighbourhood, either four or eight, nodes have four or eight edges, respectively. If setting up an irregular graph of image regions, the number of adjacent nodes and edges differs significantly depending on the image content and the applied segmentation technique.

Reconsidering nodes a_i and b_i in figure 3.2(c), node a_i receives and sends five messages whereas b_i only two. Nodes with many neighbours receive more messages and consequently gain a higher weight than nodes with less neighbours (details of training and inference in Appendix A). In addition, the impact of the association potential of a node on its label will significantly decrease the more messages are received via edges. The label of node a_i would basically become a function of its neighbouring nodes, its own features would significantly lose importance. A very high number of edges would lead to the label of that node being almost independent of its association potential. In order to avoid this bias, each edge feature vector $\boldsymbol{\mu}_{ij}(\mathbf{x})$ is normalized through the sum of norms of feature vectors $\mathbf{h}_j(\mathbf{x})$ of neighbouring nodes at that particular node to obtain $\boldsymbol{\mu}_{ij,irregular}(\mathbf{x})$:

$$\boldsymbol{\mu}_{ij,irregular}(\mathbf{x}) = \boldsymbol{\mu}_{ij}(\mathbf{x}) \,/ \sum_{j \in N_i} \|\mathbf{h}_j(\mathbf{x})\|. \tag{3.1}$$

In this way it is guarenteed that each node receives the same amount of messages via its edges during inference. No priority is given to nodes with more neighbours and all nodes have per se equal weighting. It is noteworthy that this graph of regions is anisotropic in contrast to the patch graph. The value of an edge potential between two nodes in the region graph depends on the direction the message is passed, whereas this is not the case for the isotropic regular graph of image patches. In figure 3.2(c), a message passed from h_i to a_i receives another weighting as vice versa, for example, because a_i has five neighbours and h_i only three. In the following section novel techniques to exploit the region-graph for building detection in remote sensing data will be introduced.

3.1.2. Gradient-based discontinuity constraint

One possible way to learn context in remote sensing data of urban scenes more expressively exploiting the region-graph topology is proposed. Reconsidering the standard formulation of CRFs (Eq. 2.15), two terms enable context integration: Association potential $A_i(\mathbf{x}, y_i)$ and interaction potential $I_{ij}(\mathbf{x}, y_i, y_j)$. In this section, focus is on learning context with $I_{ij}(\mathbf{x}, y_i, y_j)$, in the following section a novel concept to incorporate context into $A_i(\mathbf{x}, y_i)$ will be introduced.

Usually, feature vector $\boldsymbol{\mu}_{ij}(\mathbf{x})$ is simply calculated by either subtracting or concatenating $\mathbf{h}_i(\mathbf{x})$ and $\mathbf{h}_j(\mathbf{x})$. In general, $\boldsymbol{\mu}_{ij}(\mathbf{x})$ can also be chosen based on other features than such already used for the association potential and other methods of comparing features are possible, too. Making use of this flexibility provided by the CRF framework, a novel design of the CRF interaction potential for building detection is proposed.

There are mainly two reasons for a specific design. The standard interaction potential is very appropriate for detecting few instances of an object class in an image (Fig. 3.1(a,b)). Most times, only one instance per object class appears in an image and it often covers a rather large homogeneous area. If dealing with remote sensing data, many instances of the same object class appear in various locations within an image (Fig. 3.1(c,d)). Small gaps are located in-between buildings and the area of a single building is small compared to the image size. Additionally, building roofs and neighbouring streets or parking lots often have very similar features. Corresponding nodes are hard to discriminate into building and non-building considering only feature vectors. However, regions with similar features can belong to different object classes if they are separated by a high gradient in the optical image. Figure 3.3(b) shows a gradient magnitude image of the optical image in figure 3.3(a). High gradients often result from building roof edges, which can be seen as black lines in figure 3.3(b). The idea is to introduce the gradient between two nodes as an explicit discontinuity constraint into the interaction potential. Two nodes belong to the same class if a low gradient arises in-between, whereas they are part of different classes if separated by a high gradient.

The same idea is proposed in Wegner et al. [2011a], but on a regular grid of image patches. A drawback of image patches is that they cut across object boundaries. Often, meaningful gradients are not located between two patches, but inside a patch. Moreover, a patch grid naturally prefers horizontal and vertical gradients, which is disadvantageous if dealing with an urban scene of arbitrary

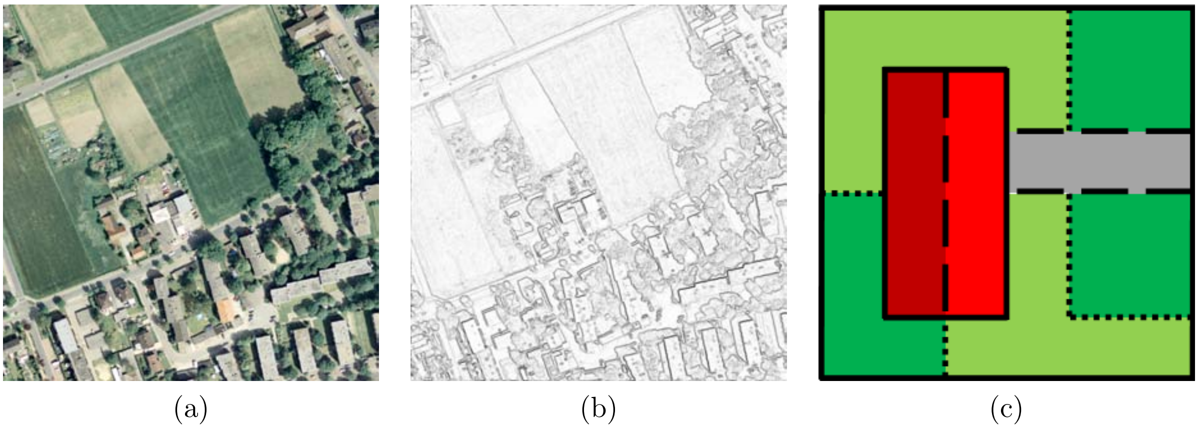


Figure 3.3.: (a) Optical aerial image, (b) gradient magnitude of its intensity channel (black: high magnitude, white: low magnitude), (c) multi-scale segmentation (three scales) into regions with gradients of three different magnitudes (solid line: coarse scale, dashed line: medium scale, dotted line: high scale).

orientation with respect to image coordinate axes. A more appropriate technique is to set up a graph on image regions, as described in the previous section, better preserving object boundaries. Segmentation algorithms aggregate homogenous pixels (with respect to a particular criterion) and separate regions if sufficient gradient occurs. The degree the gradient magnitude is considered to discriminate regions depends on the segmentation algorithm. This is valid in general for algorithms like watershed, Normalized Cuts [Shi & Malik, 2000], or Mean Shift [Comaniciu & Meer, 1999]. A particularly convenient segmentation method for the proposed approach is Quickshift [Vedaldi & Soatto, 2008], which is closely related to Mean Shift. Basically, it aggregates pixels of similar hue, saturation, and intensity that are located closely to each other. It orders regions of multiple scales in a tree and allows to choose the appropriate region sizes by cutting through the tree at a desired scale. All smaller regions of higher scales are contained within larger regions of coarser scales. At the coarsest scale an image is partitoned into only a few regions being separated by high gradients. The example scene in figure 3.3(c) is subdivided into two segments framed with solid lines, the first one containing the building, the second one street, trees, and grassland. At the next higher scale large regions are split into smaller regions separated by gradients of smaller magnitude shown with dashed lines in figure 3.3(c). The highest scale again subdivides regions of the previous scale (dotted lines in Fig. 3.3(c)) and generates those regions that are used to set up the region-graph (cf. 3.2(c)). Original boundaries of large regions generated at coarse scales are kept and gradients in-between, too. In an ideal case, a high gradient separates buildings from their environment, while low gradients occur inside buildings or the non-building class.

In order to transfer this idea to the CRF, each element in feature vector $\boldsymbol{\mu}_{ij}(\mathbf{x})$ of the edge between nodes i and j is multiplied with a scalar weight $w_{disc,ij}$. It is a function of the mean gradient magnitude g_{ij} between two regions (i.e., the gradient along the common border) scaling the corresponding feature vector. Discontinuity preserving edge feature vector $\boldsymbol{\mu}_{ij,disc}(\mathbf{x}, w_{disc,ij})$ is expressed as given in equation 3.2.

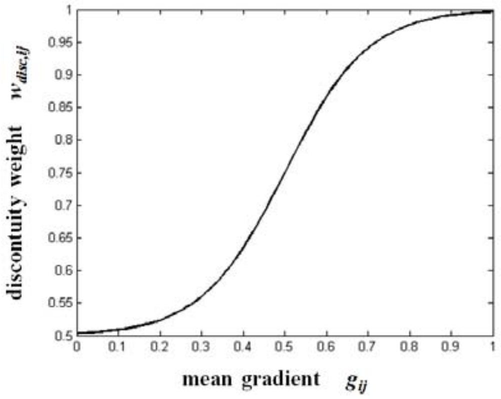


Figure 3.4.: Sigmoid discontinuity weighting function with parameters $\alpha = 10$ and $\kappa = 0.5$ (note: y-axis begins at 0.5)

$$\boldsymbol{\mu}_{ij,disc}\left(\mathbf{x}, w_{disc,ij}\right) = \boldsymbol{\mu}_{ij}\left(\mathbf{x}\right) w_{disc,ij}. \tag{3.2}$$

An investigation of the histogram of g_{ij} of the entire image suggested that, in general, values above 0.5 indicate that two regions belong to different object classes. The mean gradient g_{ij} is not used directly as weighting parameter (which would correspond to a linear weighting), but introduced into a sigmoid function[1] with the inflexion position at $\kappa = 0.5$. Various settings of α were tried resulting in an optimal value of $\alpha = 10$. Additionally, the sigmoid function is shifted in y-direction in order to still allow for discontinuities between two adjacent regions i and j if features indicate different classes without being separated by a high gradient. The resulting sigmoid weighting function is shown in figure 3.4, the corresponding equation is 3.3.

$$w_{disc,ij} = \left(1 + \left(\frac{1}{1 + \exp\left(-\alpha\left(g_{ij} - \kappa\right)\right)}\right)\right)/2 \tag{3.3}$$

The discontinuity preserving edge feature vector $\boldsymbol{\mu}_{disc,ij}\left(\mathbf{x}, w_{disc,ij}\right)$ is introduced into the linear model of the interaction potential and replaces the standard edge feature vector $\boldsymbol{\mu}_{ij}\left(\mathbf{x}\right)$:

$$I_{ij}\left(\mathbf{x}, y_i, y_j\right) = y_i y_j \mathbf{v}^T \boldsymbol{\mu}_{disc,ij}\left(\mathbf{x}, w_{disc,ij}\right). \tag{3.4}$$

Elements of the original edge feature vector $\boldsymbol{\mu}_{ij}\left(\mathbf{x}\right)$ that is scaled based on the gradient can be generated in different ways. One possiblity are bounded ratios of node feature vectors [Wegner et al., 2011a]. Optimal results are achieved using the absolute differences as shown in section 2.2.4. Basically, the gradient-based discontinuity weighting suppresses or supports smoothing. A more sophisticated formulation that turns away from simple smoothing terms is presented in the following section.

[1] Sigmoid functions are common for weighting if a compromise between linear weighting and step function is necessary for a specific application (e.g., [Tupin & Roux, 2005]).

3.1.3. Implicit scene context

In the previous section it was shown how contextual knowledge can be modelled via the interaction potential $I_{ij}(\mathbf{x}, y_i, y_j)$ (right term in Eq. 2.15). Another possibility is to model context with the association potential $A_i(\mathbf{x}, y_i)$ (left term in Eq. 2.15). In this section a novel method is proposed integrating data globally, thus exploiting the definition of $A_i(\mathbf{x}, y_i)$ to its full extent. Even though computing features in several resolutions (cf. section 2.2.4) enlarges a specific local neighbourhood beyond the capabilities of MRFs, the methods presented here in sections 3.1.1 and 3.1.2 as well as most of the techniques reviewed in section 1.3.2 (e.g., [He et al., 2004; Kumar & Hebert, 2006]) rest quite local. The definition of CRFs allows to consider all data $\mathbf{x}$ in association and interaction potential, no restrictions exist with respect to location or correlation of features.

The key idea of this approach is to capture context of the background class, the so-called *implicit scene context* (ISC), of partially labeled images via histograms to support object segmentation and classification [Wegner et al., 2011b]. With partially labeled it is meant that only a small portion of object categories existing in data are semantically annotated in training data. In this way object-context in an image can be used for classification without giving semantics to each object explicitly. Classification is limited to a binary decision between an object we want to detect and all other object categories in the background class. All categories not explicitly labeled are contained within a joint background class. Object classes may contain characteristic subcategories, too. Figure 3.5 demonstrates what is meant by background class, object class, and subcategories. In figure 3.5(a) the two classes building and non-building to be discriminated are differently shaded. Class non-building consists of three subcategories: grassland (vertically shaded light green), trees (horizontally shaded dark green), and street (screened light grey). The building class contains two subcategories (Fig. 3.5(c)): one side of a gable roof facing the sun (shaded light red) and the opposite shadowed roof plane (shaded dark red). These subcategories of two classes building and non-building are characteristic for urban areas and they often occur in particular patterns (Fig. 3.5(d)). Driveways connect streets with buildings, buildings are surrounded by grassland and trees (cf. 3.1(c,d)), for example. The aim is to capture and learn these patterns without having to label all subcategories explicitly in the training database for several reasons:

- Labeling only two classes building and non-building (i.e., background) saves time,
- a binary classification task reduces the amount of features and training data needed compared to a multi-class setup,
- one does not explicitly have to know all object classes contained in the data beforehand,
- and the context level of detail can be chosen by a parameter of the algorithm instead of having to label all training data again if a more detailed context level is required.

The following requirements have to be met by the algorithm: It should be able to cope with very local to global context scales. In addition, ISC shall be kept generically applicable to multiple kinds of scenes. It should capture, for instance, context in terrestrial images of building facades,

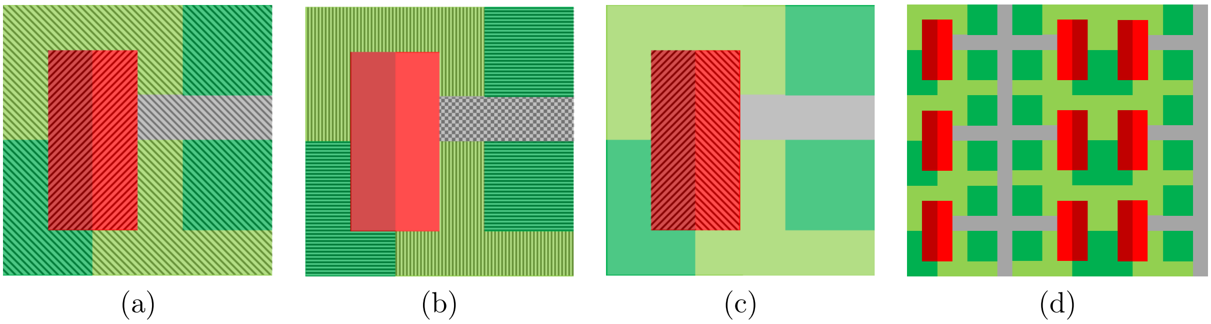


Figure 3.5.: Partially labeled example image: (a) two classes building and non-building, (b) implicit object categories contained in non-building (background) class, (c) implicit categories contained in building class, (d) global pattern of implicit scene context.

where usually sky is above the facade and vegetation below, but also in aerial images of buildings, where no preferred ordering with attributes like "above" and "below" exists. Thus, no preferred direction should be relied on. Finally, computational efficiency shall be achieved and computation of co-occurrences be avoided.

Inspired by the "thing and stuff" (TAS) concept of Heitz & Koller [2008] and the "shape context" histograms of Belongie et al. [2002], implicit scene context is proposed to augment CRFs (ISC-CRF). A general formulation able of capturing background context and its relation to object classes via histograms is introduced. Integration into a CRF is possible without major changes to the general framework in terms of training and inference, which is an important difference to the works of Kohli et al. [2009] who add a third term, the region consistency potential, to the traditional association and interaction potentials of pair-wise CRFs. Neither an additional potential is added nor any complex graph structure is generated, but the flexibility provided by the definition of the association potential, which depends on all data globally [Kumar & Hebert, 2006], is exploited. This technique allows for very local up to global contextual learning. Computationally expensive co-occurrence statistics of object categories [Rabinovich et al., 2007; Ladicky et al., 2010] are avoided by representing context via histograms as done by Belongie et al. [2002], Wolf & Bileschi [2006], and Savarese et al. [2006]. An ISC-CRF has the following properties:

- Characteristic patterns within the background class of partially labeled images and their relation to labeled object classes are learned.
- Contextual patterns are formulated in terms of histograms. Rotation invariance is achieved and the use of multiple context scales ensures good performance for both small and big objects.
- Although it is modelled as a unary potential within a CRF framework, it can generally be utilized (with minor changes) with any kind of non-contextual classifier like Support Vector Machines, too.

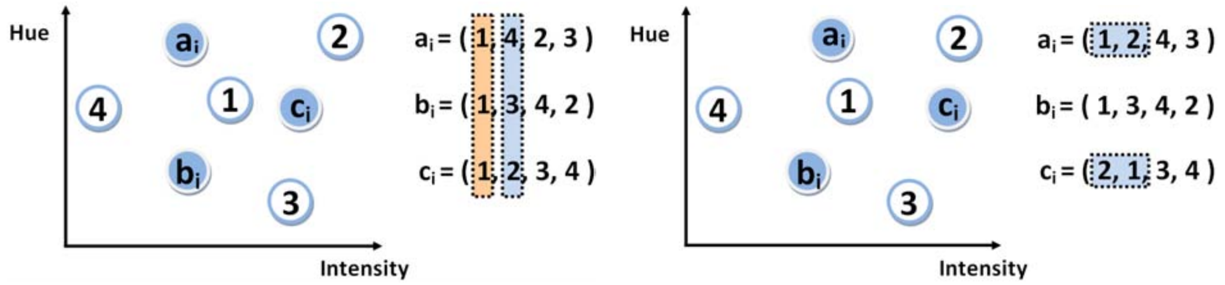
This novel approach is generally applicable to arbitrary image scenes, for example, aerial, terrestrial, and medical images. We can benefit from very large databases of only partially labeled images and learn context although we do not explicitly know all object classes. In addition to the object classes that have been explicitly labeled for training, we can use patterns existing in the unlabeled part of the data (i.e., labeled as background class). All steps necessary for training will be explained next followed by a description of the testing phase. In order to meet the requirements aforementioned, training consists of:

- Multi-scale segmentation of images into regions,
- computation of features per region in all scales,
- unsupervised k-means clustering based on the previously generated features,
- generation of implicit context histograms in three different ranges per region,
- computation of histogram features,
- integration as feature vector into the CRF unary potentials,
- and training of the CRF based on labeled images.

It should be noticed that labels of training data are considered for the first time during the last step, the CRF training. All processing before (multi-scale segmentation to integration into CRF unary potentials) is done without assigning labels building or non-building to the regions. Solely the training images (not the two label categories) are used until CRF parameter adjustment in order to capture subcategories of both classes. Building class and non-building class contain several previously unknown subcategories. In an ideal case, the cluster centers of k-means describe building and non-building subcategories, one center for each. This is convenient because different building types occur, big flat roof buildings and small gable roof buildings, for example. Each of these types is embedded into a characteristic context, too, which is captured via specific context histograms. Assigning labels explicitly right at the beginning would lead to background context being learned only for a single building class. Moreover, this background context would be less specific because context of small gable roof and big flat roof buildings would be mixed within one feature. As a consequence, introducing binary labels at the last training step enables a very comprehensive scene description because the variability of buildings and their corresponding typical environment is learned.

An unsupervised classification of all regions is performed first for training. Any kind of unsupervised classifier could be applied, but for means of speed and simplicity a standard k-means clustering is chosen. As input to k-means clustering all features $\mathbf{h}_i(\mathbf{x}) \in \mathbf{h}(\mathbf{x})$ computed per region are taken. The cluster centers $\mathbf{K}$ generated with k-means clustering $\mathbf{K} = K_{means}(\mathbf{h}(\mathbf{x}))$ are used for the following processing.

After k-means clustering, distances to all cluster centers $\mathbf{K}$ are determined in feature space for each region. Cluster indices $\mathbf{y}_{us}$ are recorded in ascending order in a vector per region according to their distances, the closest center first, the furthest last. Recording not only the closest center, which would correspond to a Minimum Distance classifier, but all others in ascending order, too, has advantages in terms of descriptive context learning and robustness.



(a) (b)

Figure 3.6.: Two-dimensional feature space spannend by hue and intensity: Nodes a_i, b_i, and c_i and cluster centers 1, 2, 3, and 4 are shown; cluster centers are recorded in descending order with respect to their distances to the nodes; (a) nodes cannot be distinguished based on closest cluster center (first vector elements in orange frame), but on the second closest (second vector elements in blue frame), (b) gain in robustness: although the closest cluster centers of nodes a_i and c_i are different, they belong to the same class because any combination of the first two vector elements (framed in blue), no matter their order, is learned to be descriptive.

Figure 3.6(a) shows an example consisting of three nodes a_i, b_i, and c_i in blue circles with white frames in feature space defined by hue and intensity. Cluster centers 1 to 4 computed with k-means (considering additional nodes to the ones shown in Fig. 3.6) are depicted in white circles with blue frames. Indices 1 to 4 are the indices of the cluster centers, the vector of all indices is $\mathbf{y}_{us}$. Assuming a_i and c_i to belong to building subcategories and b_i to a non-building subcategory, it would be impossible to distinguish them if taking merely the closest cluster index because all three nodes have equal distances to cluster center one. If just recording the closest center (first element in vectors in Fig. 3.6(a) framed in orange), all nodes would be labeled one, although they occur at different positions in feature space. The second closest cluster center (framed in blue) is different for all nodes and helps distinguishing.

In order to explain the gain in robustness, figure 3.6(b) shows a slightly different setup. Nodes a_i and c_i, sharing the same class, have distinct closest cluster centers. Nonetheless, considering in addition the second closest elements, too, both nodes share the same first two cluster centers (framed in blue), only their order changes. A feature is defined that accounts for this varation of absolute ordering. Regions are considered to be located closely in feature space if the first two vector elements are equal, no matter their order. In conclusion, benefits are twofold: First, the type of cluster centers at each node carries valuable information facilitating detailed distinctions between classes, second, robustness is gained if nodes of the same class are assigned to equal cluster centers, but in different orders.

An example of resulting labeled regions of the closest cluster centers are shown schematically in Fig. 3.7(a). Five distinct subcategories occur, captured with $k = 5$ cluster centers[2]. For means of

[2] The number of cluster centers has to be set manually a priori. Experimental results with varying cluster center numbers (4.2.4) indicate that choosing more centers than subcategories contained in data does not significantly deteriorate performance. Automatic determination of the exact number of subcategories in feature space, based

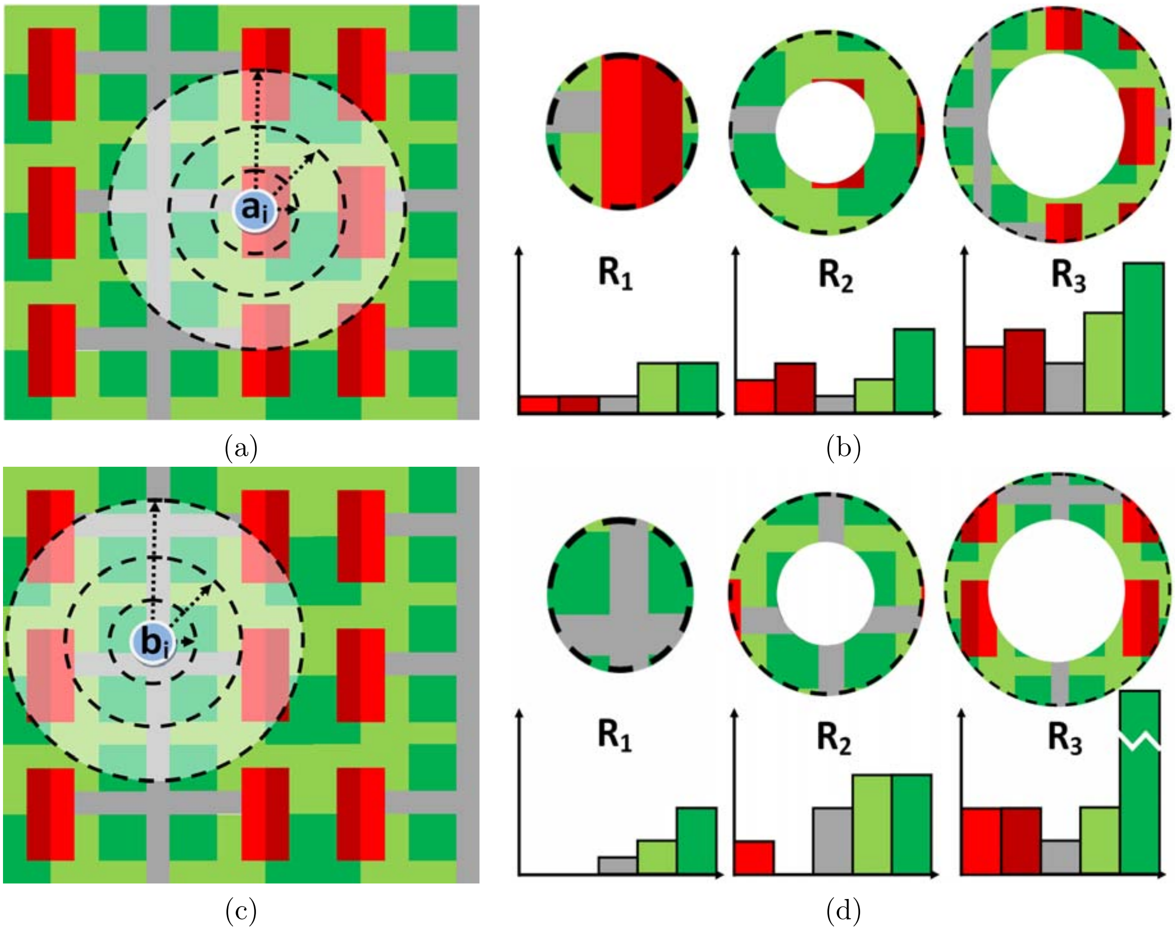


Figure 3.7.: Principle of implicit context: (a, c) ranges around the centroid of a region belonging to subcategory "light red roof" (part of building class) represented by node a_i and a region belonging to subcategory "street" (part of non-building class) represented by node b_i, (b, d) histograms of cluster labels of three ranges $R1$, $R2$, and $R3$; the ordinate counts the number of regions per cluster label within a range R, cluster labels are ordered on the abscissa; colours indicate different cluster labels appointed to regions; region boundaries run along colour edges.

understandability, only the indices of the closest cluster centers (i.e., first elements of vectors in Fig. 3.6) are depicted, second closest centers etc. are not shown. Next, the centroid C_S of each region is determined and histograms of labels $hist_R(\mathbf{y}_{us})$ occurring within different ranges R around each region are generated. Numbers of label occurrences $\mathbf{y}_{us}$ within a range R are counted in histograms. The way this is done is shown in figures 3.7(a,b) for a node a_i of subcategory "light red roof" and in figures 3.7(c,d) for a node b_i of subcategory "street". Occurrences of the five different labels are counted in three ranges $R1$, $R2$, and $R3$[3]. This procedure is conducted for all nodes in the graph. In figure 3.7(c) node b_i represents subcategory "street" (which is part of the background). Again, label occurrences within the same three ranges are counted and stored in histograms (Fig.

on the ISODATA method [Ball & Hall, 1967], for example, is left for future work.

[3] Any number of ranges can be chosen depending on the scene and on the scale of context. However, more ranges lead to increasing computational costs; three ranges are usually sufficient.

3.7)(d). Those histograms show distinct shapes in all ranges for different subcategories (cf. Fig. 3.7(b) & (d)). They capture the characteristic environment of each subcategory as a function of distance. Combining histograms of all ranges ($R1$, $R2$, and $R2$ in Fig. 3.7) results in distinct context distributions of all subcategories. It should be reconsidered that no labels of the two classes building and non-building have explictly been assigned to any node, yet. Either short or long ranges can be chosen depending on whether local or global context is to be integrated. It should be noted that longer ranges do not lead to any more complex graph structure because no graph is set up at this point at all. Furthermore, the number of ranges and either coarse or fine scaling facilitates to capture the distribution of object categories contained in the background class as a function of their distance to the node of interest. In order to meet the requirements of generalizability and transferability to multiple object classes and scenes, the exact ranges should be adapted to the scale of the context. The scale of the desired object class and its context can be approximated via the size of image regions after (over-)segmentation. Ranges R as a linear function of the mean region size were found to be optimal after tests with different image data and scenes.

Various moments and additional information representing contextual patterns in the environment of a particular region are derived from the histograms. It is noteworthy that label histograms can either be directly introduced to node feature vectors or specific features can be derived from histograms, the index of the most often appearing label within each range, the index of the label covering the largest area, for example. Qualitative, quantitative, and spatial context features $\mathbf{C}\left(\mathbf{h}\left(\mathbf{x}\right)\right)$ may be generated.

For the testing phase, exactly the same processing steps are applied except k-means clustering (and CRF training). Those cluster centers $\mathbf{K}$, originally generated with k-means during training, are used to determine closest cluster centers in ascending order per region of test data. Cluster indices are determined for all test data nodes (i.e., regions of the test images after segmentation), measuring distances in feature space to cluster centers generated in the training phase. Again, not only the closest cluster center is recorded, but all of $\mathbf{K}$ (cf. Fig. 3.6). Context histograms of several ranges are determined in the test images capturing distinct context distributions as function of distance per subcategory. Implicit context features $\mathbf{C}_i\left(\mathbf{h}\left(\mathbf{x}\right)\right)$, both of training and test data[4], are introduced into the standard linear model of the association potential as described in section 2.2.4:

$$A_i\left(\mathbf{x}, y_i\right) = y_i \mathbf{w}^T \mathbf{C}_i\left(\mathbf{h}\left(\mathbf{x}\right)\right) \tag{3.5}$$

The class of each region i can be derived merely based on implicit context features $\mathbf{C}_i\left(\mathbf{h}\left(\mathbf{x}\right)\right)$ or local node features $\mathbf{h}_i\left(\mathbf{x}\right)$ can be added to the feature vector, too. Pair-wise potentials only change in such a way that the element-wise absolute differences between nodes i and j in the graph are computed based on the corresponding implicit context features (Eq. 3.6).

[4] For the sake of clarity: It is not done simultaneously, but first during training. The log-likelihood objective function of equation 2.15 is derived and parameters are tuned within an optimization framework (details in Appendix A). Thereafter, testing is carried out on new unlabeled data and inference takes place using the parameters that were tuned during training.

$$I_{ij}\left(\mathbf{x}, y_i, y_j\right) = y_i y_j \mathbf{v}^T \boldsymbol{\mu}_{\mathbf{C},ij}\left(\mathbf{x}\right), \quad \boldsymbol{\mu}_{\mathbf{C},ij}\left(\mathbf{x}\right) = \left|\mathbf{C}_i\left(\mathbf{h}\left(\mathbf{x}\right)\right) - \mathbf{C}_j\left(\mathbf{h}\left(\mathbf{x}\right)\right)\right|. \tag{3.6}$$

No normalization of the label count in the histogram is done based on the size of the regions, for example, because tests show that the importance of a region does not necessarily increase with its size. Small regions can be characteristic context features and are of high relevance for a particular object class, too. Dealing with a multi-scale segmentation, implicit context histograms can be computed at coarser scales, too. It is possible to learn global context of coarse scene structures at a coarse scale while simultaneously capturing local context at the finest scale[5].

3.2. Accuracy assessment of building height estimation

A novel approach to building height estimation combining measures of one SAR acquisition and an optical image within a least squares adjustment framework is proposed in this section. The objective is to investigate the theoretical accuracy that can be achieved. It is shown how optical and SAR effects as well as the different sensor viewing geometries of both sensor types can be used to estimate heights of buildings. Emphasis is on geometric effects rather than radiometric ones.

Much research has been conducted estimating shape and height of buildings via a detailed modelling of their electromagnetic properties [Guida et al., 2008, 2010], which have to be known a priori. In real world scenarios we only partially have access to the electromagnetic properties of different parts of a building and often we do not have such prior information at all. Another possibility is to apply a simulation and matching procedure for height estimation as done by Brunner et al. [2010]. Their strategy implies that a three-dimensional building model, needed for simulation, has to be manually generated based on an optical image. Simulation and matching only succeeds if no signal interferences due to adjacent objects occur in the original SAR image. Simulation resulting in restrictions on the scene content shall be avoided here. The aim in this section is to directly use the height information that is contained in optical imagery, SAR, or InSAR data, and their combination.

Different possibilities of height measurements exist, some were already mentioned in section 2.1. Each measurement provides another height value for the same building. This measurement can either be based on a single data source or on a combination of both. Those different height measurements are combined within a least squares adjustment framework (Gauß-Helmert-Modell). One stochastically sound height estimate per building and a corresponding precision are computed. Not only a single height estimate is needed, but a precision, too, in order to judge its quality. Thus, single SAR acquisition and optical image are combined in two ways: First, for height measurements and, second, in the least squares adjustment framework. In the following subsection different possibilities to measure heights of buildings with flat and with gable roofs are proposed.

[5] Graphs of image regions generated with a multi-scale segmentation can also be used directly for classification if object shapes are learned via so-called region ancestries as proposed by Lim et al. [2009]. The integration of this promising concept into a CRF framework is left for future work.

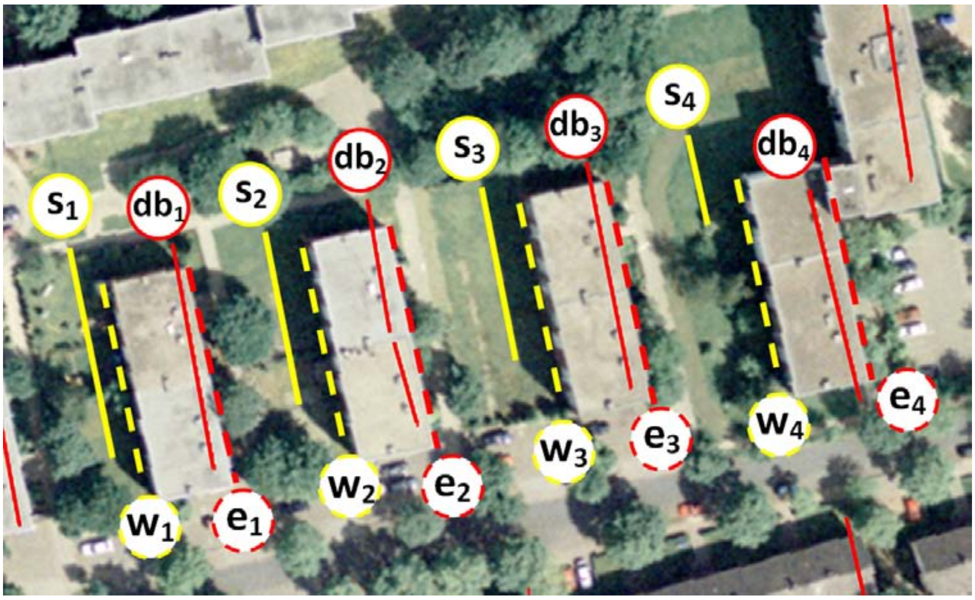


Figure 3.8.: Roof edges e, shadow edges s, double-bounce lines db, and wall meeting ground at w overlaid to a cut-out of the optical image containing flat roof buildings.

3.2.1. Height measurements

Several methods to measure building heights in SAR and optical data exist. Considering InSAR, heights can directly be obtained after phase to height conversion. Interferometry is capable of providing highly accurate height information, but phase unwrapping is often hard to solve particularly in urban areas because of sudden height jumps at buildings. If a building is higher than the maximum unambiguous height Δh (cf. 2.3), its altitude can hardly be determined. Additionally, SAR phenomena like layover, shadowing, and interfering backscatter of multiple objects in the same resolution cell complicate automatic analysis. Nonetheless, they contain valuable information about the geometry of the object under investigation, too. New sources of building height information extraction open up if combining effects of SAR and optical sensor.

Considering a single optical image, height information is contained in the shadow of the sun and geometrical distortion caused by the central perspective of the camera[6]. Knowing the azimuth of the sun, either given by a timetable or measured in the image, building height h_s is a function of sun incidence angle ρ, the location $\boldsymbol{w}$ where building walls in the optical image meet ground, and shadow edge position $\boldsymbol{s}$ (Eq. 3.7 & Fig. 3.9). Distance $d_s = \|\boldsymbol{s} - \boldsymbol{w}\|$ between w and s has to be measured parallely to the sun rays.

$$h_s(\rho, \left[\begin{smallmatrix} w_x \\ w_y \end{smallmatrix}\right], \left[\begin{smallmatrix} s_x \\ s_y \end{smallmatrix}\right]) = \tan(\rho) d_s = \tan(\rho)\sqrt{(s_x - w_x)^2 + (s_y - w_y)^2}. \tag{3.7}$$

[6] Line scanners, as optical satellite sensors, only have a central perspective orthogonally to their direction of flight. For means of simplicity only standard frame cameras with a central perspective, shown schematically in figure 3.9, are considered here.

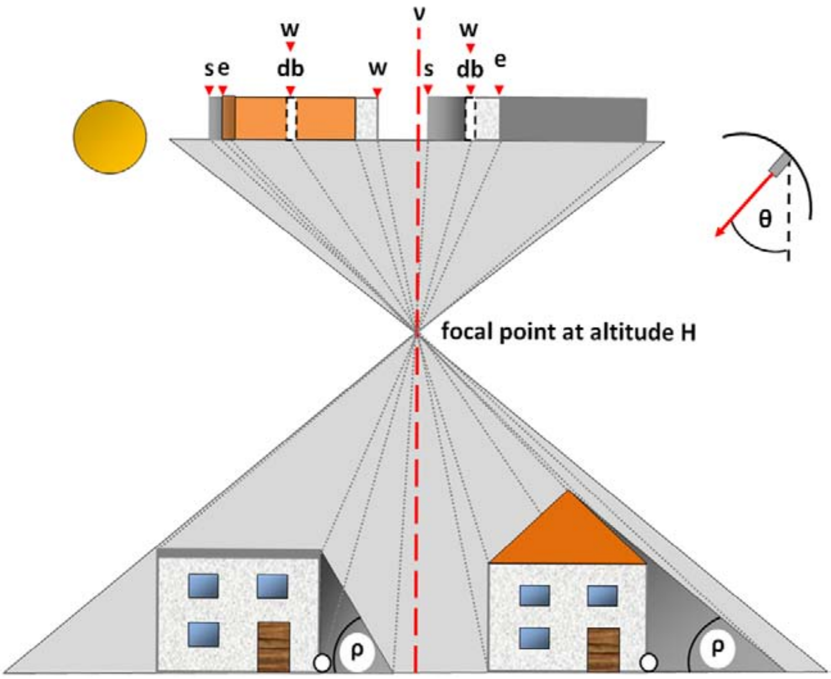


Figure 3.9.: Height measurement with an optical image and double-bounce lines derived from SAR data: Sketch of trigonometric relations of parameters; location of double-bounce lines is shown with white circles where building walls meet ground (ν: nadir, db: double-bounce line projected to the optical image, e: building edge, s: shadow boundary, ρ: sun incidence angle, θ: SAR sensor viewing angle, w: edge where building wall meets ground).

Figure 3.8 shows positions of s and w at four flat roof buildings in a cut-out of an optical orthophoto. The position of w is usually hard to determine due to low contrast between ground and facade, which are shadowed. Dashed yellow lines depict positions of w in figure 3.8. Such measurement only works if assuming vertical building walls, locally flat terrain next to the builiding, and no obstruction by adjacent objects. Only a small part of shadow edge s_4 can be used for measurements because it is occluded by trees. Shadow edges s_2 and s_3 as well as corresponding w_2 and w_3 satisfy all assumptions, height measurements are possible. Shorter shadows (i.e., greater sun incidence angles ρ and smaller builidings) decrease the height accuracy being limited by measurement precisions of w and s (as function of ground sampling distance). A schematic sketch of sun incidence angle ρ and mapping of s is given in figure 3.9. The position of w coincides with the place in the image where db is mapped, too. In case sun and nadir of the optical sensor (ν) are located on the same side of a building, w can hardly be determined. Position w of the gable roof building in figure 3.9, coinciding with double-bounce line position db highlighted with a white circle, is not mapped. It is overlaid by the building roof due to perspective distortion, but can be determined if a SAR double-bounce line is present (cf. position of db in the image profile top left in Fig. 3.9).

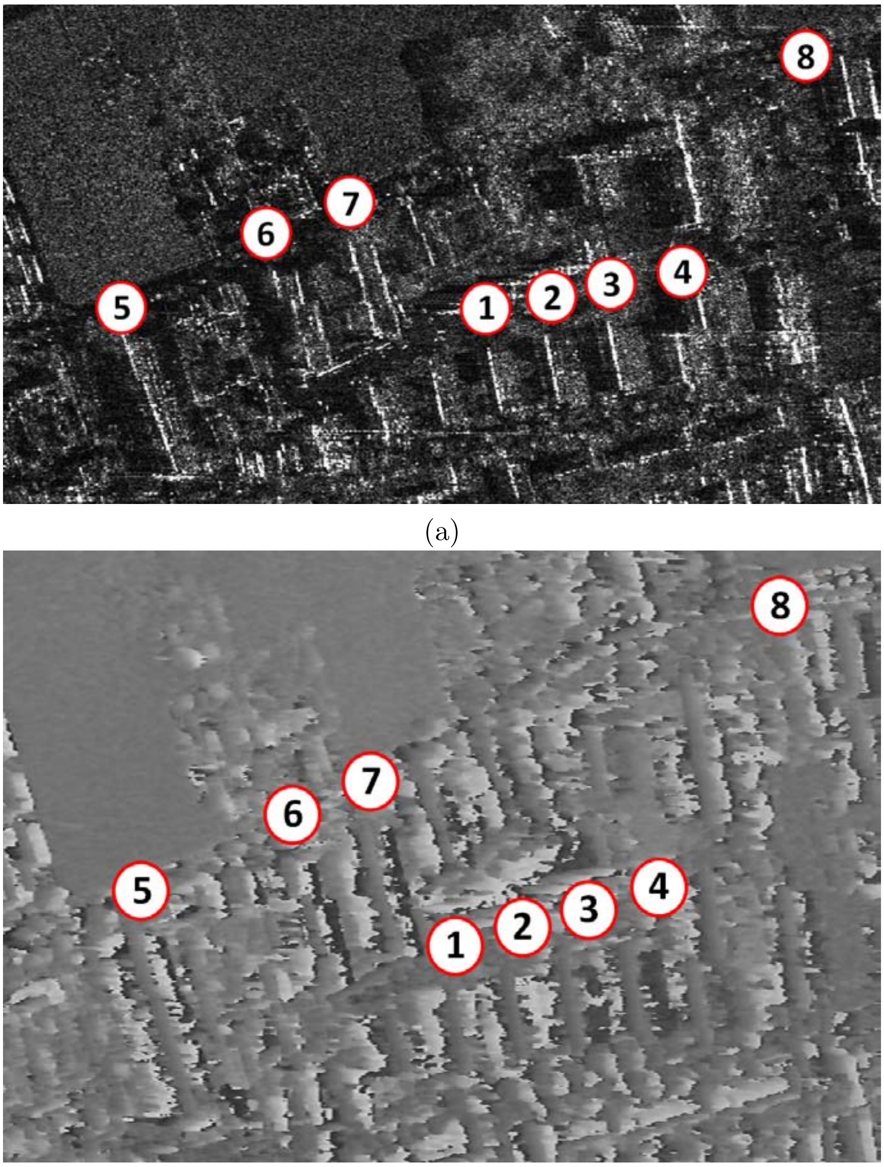
Heights of buildings can be measured combining double-bounce line and optical perspective distortion, too. Double-bounce lines are part of the building footprint (cf. 2.1.4). Their position corresponds to w where building walls and ground meet (cf. Fig. 2.7(c)), highlighted with white circles in figure 3.9. Extracted double-bounce lines db are projected to the geometry of the optical image using the InSAR heights (and all necessary sensor parameters). In figure 3.8 line db is shown as solid red line at four flat roof buildings in orthophoto geometry. A SAR image and corresponding InSAR heights containing the same four buildings (1, 2, 3, 4) is given in figure 3.10, double-bounce lines appear as straight white lines in figure 3.10(a). They are overlapped by parts of roofs in the optical image due to perspective distortion (Fig. 3.8). Reconsidering that db, being part of the footprint, corresponds to w, height h_{db} depends on overlap of roof edge e (depicted as dashed red line in Fig. 3.9) and SAR double-bounce line db. It depends on relation $\|[\begin{smallmatrix} e_x \\ e_y \end{smallmatrix}]-[\begin{smallmatrix} \nu_x \\ \nu_y \end{smallmatrix}]\|/\|\begin{bmatrix} db_x \\ db_y \end{bmatrix}-[\begin{smallmatrix} \nu_x \\ \nu_y \end{smallmatrix}]\|$, scaled by sensor altitude H. More precisely, it is a function of image coordinates $[\begin{smallmatrix} e_x \\ e_y \end{smallmatrix}]$ of points on roof edge e, double-bounce line $\begin{bmatrix} db_x \\ db_y \end{bmatrix}$, nadir point $[\begin{smallmatrix} \nu_x \\ \nu_y \end{smallmatrix}]$, and optical sensor altitude H:

$$h_{db}\left(\begin{bmatrix} db_x \\ db_y \end{bmatrix}, [\begin{smallmatrix} e_x \\ e_y \end{smallmatrix}], [\begin{smallmatrix} \nu_x \\ \nu_y \end{smallmatrix}], H\right) = H \cdot \left(1 - \sqrt{\frac{(db_x - \nu_x)^2 + (db_y - \nu_y)^2}{(e_x - \nu_x)^2 + (e_y - \nu_y)^2}}\right). \tag{3.8}$$

The distance between db and e is measured orthogonally to e. Line db_2 in figure 3.8 is split into two parts, not exactly parallel to roof edge e_2, because the bright line of building 2 in the SAR image in figure 3.10(a) has a low signal return directly at the gap. An investigation of the optical image (Fig. 3.8) suggests a tree right in front of the building facade as reason for this disturbance. An inaccuracy of the InSAR height values, possibly due to mixed signal return of facade and tree, leads to small displacements if lines are projected to orthophoto geometry (cf. Fig. 3.10(b)). Both parts of db_2 consequently are not exactly parallel to roof edge e_2. This artefact is accounted for by computing the mean orthogonal (with respect to e_2) distance between db_2 and e_2.

Smaller heights and shorter distances to the nadir point ν (left of buildings in Fig. 3.8) lead to less perspective distortions in the optical image. Distances between double-bounce line db and roof edge e decrease leading to inaccurate distance measurements, which are limited by double-bounce line positioning accuracy and pixel measurement precision in the image. If buildings are located directly in nadir position of the optical sensor, roof edge e and double-bounce line db overlay completely. Distances $\|[\begin{smallmatrix} e_x \\ e_y \end{smallmatrix}] - [\begin{smallmatrix} \nu_x \\ \nu_y \end{smallmatrix}]\|$ and $\|\begin{bmatrix} db_x \\ db_y \end{bmatrix} - [\begin{smallmatrix} \nu_x \\ \nu_y \end{smallmatrix}]\|$ will have equal length, the square root in equation 3.8 will be one, and h_{db} will be zero, making a height measurement impossible.

Equation 3.8 can also be used to measure a building height h_{pd} completely relying on perspective distortion in optical data. Perspective projection according to camera orientation, the so-called central perspective, results in elevated objects being mapped slightly displaced in the image. For example, roof boundaries are not mapped directly onto the building footprint (cf. Fig. 3.8), but shifted away from nadir. This effect carries height information. If no double-bounce line db occurs where building wall meets ground, but the exact position of w can be recognized in the optical image, one can directly perform the same measure. It follows exactly the concepts (based on optical perspective distortion) as previously described, but without need for a double-bounce line. In figure 3.9 the position of w at the flat roof building is directly mapped by the optical sensor. If db would



(a)

(b)

Figure 3.10.: (a) Cut-out of a SAR image of an InSAR pair, (b) corresponding InSAR heights (bright values correspond to greater heights); Intermap Aes-1 sensor at Dorsten, Germany (range direction right to left); original slant range data flipped and scaled to approximate ground range geometry (for visualization).

be absent, one could nonetheless recognize w directly in the image (cf. position of w right in image profile) and exploit it for measuring a building height. Considering the gable roof building right in figure 3.9, h_{db} of the eave can be determined with the double-bounce line at the right side of the building, whereas h_{pd} can be measured with w on the left side (cf. left image profile).

The layover effect (cf. 2.1.2 and 2.1.4) can be used to measure a building height in a single SAR image, too. Increasing building height and viewing angle θ lead to greater layover. Under the assumption of flat signal wavefronts, several possibilities exist to determine the height of a build-

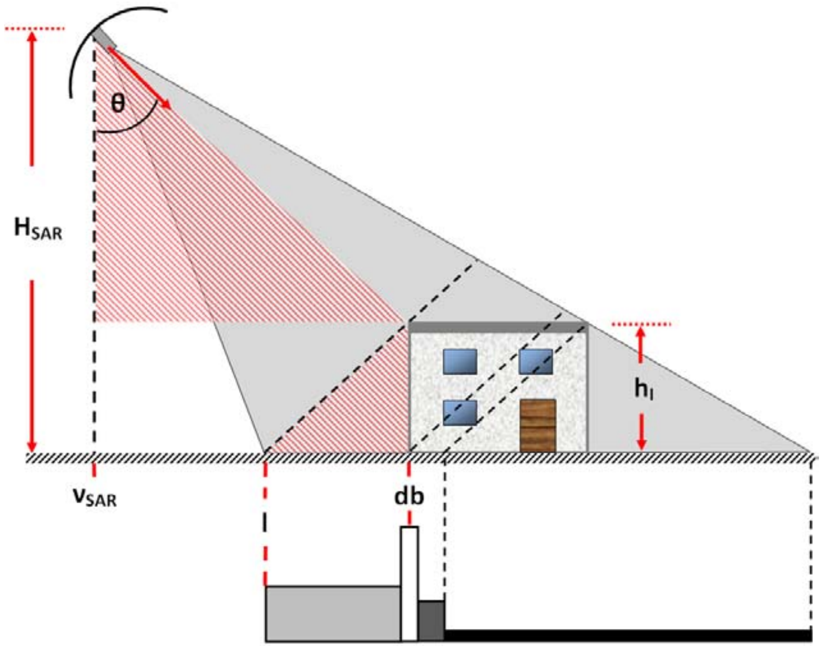


Figure 3.11.: Height measurement based on SAR layover in ground range geometry; shaded red triangles demonstrate the intercept theorem (h_l: building height, ν_{SAR}: nadir of the SAR sensor, db: double-bounce line, l: near range layover end, H_{SAR}: sensor altitude, θ: SAR sensor viewing angle).

ing based on trigonometric considerations. If a SAR image is given in slant range geometry, the height h_l can be expressed as the quotient of layover width l_{width} and cosine of viewing angle θ: $h_l = l_{width}/\cos(\theta)$. Another expression for h_l considering a SAR image in ground range geometry, avoiding the use of viewing angle θ, is found via the intercept theorem (Eq. 3.9), which is schematically shown in figure 3.11.

$$\frac{h_l}{(db - \nu_{SAR})} = \frac{(db - l)}{(H_{SAR} - h_l)} \tag{3.9}$$

Rearranging equation 3.9 we get $h_l^2 - h_l H_{SAR} + (db - l)(db - \nu_{SAR}) = 0$. To solve this quadratic polynomial for h_l, we have to complete the square delivering $h_{l;1,2} = \frac{H_{SAR}}{2} \pm \sqrt{\frac{H_{SAR}^2}{4} - (db - l)(db - \nu_{SAR})}$. The ambiguity between $h_{l;1}$ and $h_{l;2}$ can easily be solved considering that a building will never be higher than half the sensor altitude $\frac{H_{SAR}}{2}$. For example, assuming a viewing angle of $\theta = 45^\circ$ (cf. Fig. 3.11), typical parameters would be: Sensor altitude $H_{SAR} = 3000\ m$, distance between double-bounce line and nadir $db - \nu_{SAR} = 3000\ m$, and a layover width of $db - l = 20\ m$. Inserting these parameters and solving for $h_{l;1,2}$ we get $h_{l;1} = 1500\ m - 1480\ m = 20\ m$ and $h_{l;2} = 1500\ m + 1480\ m = 2980\ m$. Clearly, the solution for $h_{l;2}$ is invalid. Therefore, building height h_l can be formulated as:

$$h_l(db, l, \nu_{SAR}, H_{SAR}) = \frac{H_{SAR}}{2} - \sqrt{\frac{H_{SAR}^2}{4} - (db - l)(db - \nu_{SAR})}. \tag{3.10}$$

Dealing with InSAR data, a building height h_{InSAR} can be determined directly from the InSAR height values (Fig. 3.10(b)) contained in the layover ramp of a building (cf. Fig. 2.5(b)). First, double-bounce lines are extended to parallelograms in the InSAR height data in slant range geometry. A double-bounce line indicates the far range end of a layover ramp, its near range end is approximated by a parallel line. Both lines are completed to a parallelogram with two lines parallel to range direction. A parallelogram contains an entire layover ramp and acts as bounding box. The distance between double-bounce line and near range end of the layover ramp (width of a parallelogram) can be approximated knowing SAR sensor parameters and a rough estimate of the maximum building height in a scene. A robust maximum is determined inside each parallelogram bounding box by rank filtering: Height values are ordered in ascending order, the maximum five percent are cut off, and the remaining maximum value is taken as building height h_{InSAR}. Reconsidering the rather noisy InSAR heights shown in figure 3.10(b), h_{InSAR} is only a rough estimate with low accuracy leading to a low weight on this measurement in the least squares adjustment framework. More sophisticated methods for direct building height determination from InSAR data in urban areas, including adapted phase unwrapping techniques, have to be developed.

Radar shadow contains height information, too. In SAR images of urban areas this occlusion is often hardly visible because signal of adjacent objects interferes. If layover of a building located behind the building of interest (in range direction of the SAR sensor) falls into the shadow area, the shadow outlines cannot be detected. The area of no signal return is filled with signal return from another object, which can be observed in figure 3.10(a). Flat roof buildings one to four are characterized by bright double-bounce lines and a layover area right of it. Dark areas corresponding their radar shadows occur left of the double-bounce lines, but their left end cannot be determined precisely. Layover of neighbouring buildings and signal of trees disturbes radar shadow. In the optical image in figure 3.8 trees can be observed between buildings one and two. Considering the same area between buildings one and two in the same SAR image in figure 3.10(a), signal return of those trees falls into the radar shadow of building two. In addition, the far range end of the radar shadow of building two cannot be recognized because layover due to signal return from building one overlaps. Thus, the near range end of the layover area of building one can be observed, but not the far range end of building two's shadow.

Moreover, radar shadow may sometimes be hard to distinguish from surfaces that reflect almost no signal back to the SAR sensor. For example, very smooth (with respect to the signal wavelength) road surfaces or water bodies lead to dark areas in the magnitude image. In case a building is located next to a street, which is ubiquitous in urban areas, radar shadow of building and street can hardly be discriminated. This effect can be observed at building five in figure 3.10(a). No decision is possible whether we deal with radar shadow or street left of the double-bounce line. The same situation arises left of gable roof building six. For these reasons, radar shadow is not used for building height measurements in this thesis because the focus is on urban areas.

The previously explained equations work well if dealing with flat roof buildings. In case of gable roof buildings, some basic assumptions have to be reconsidered. For example, a building height

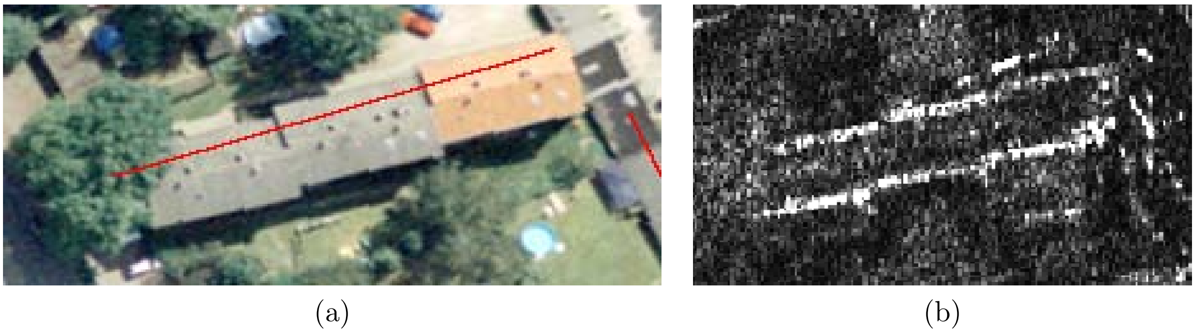
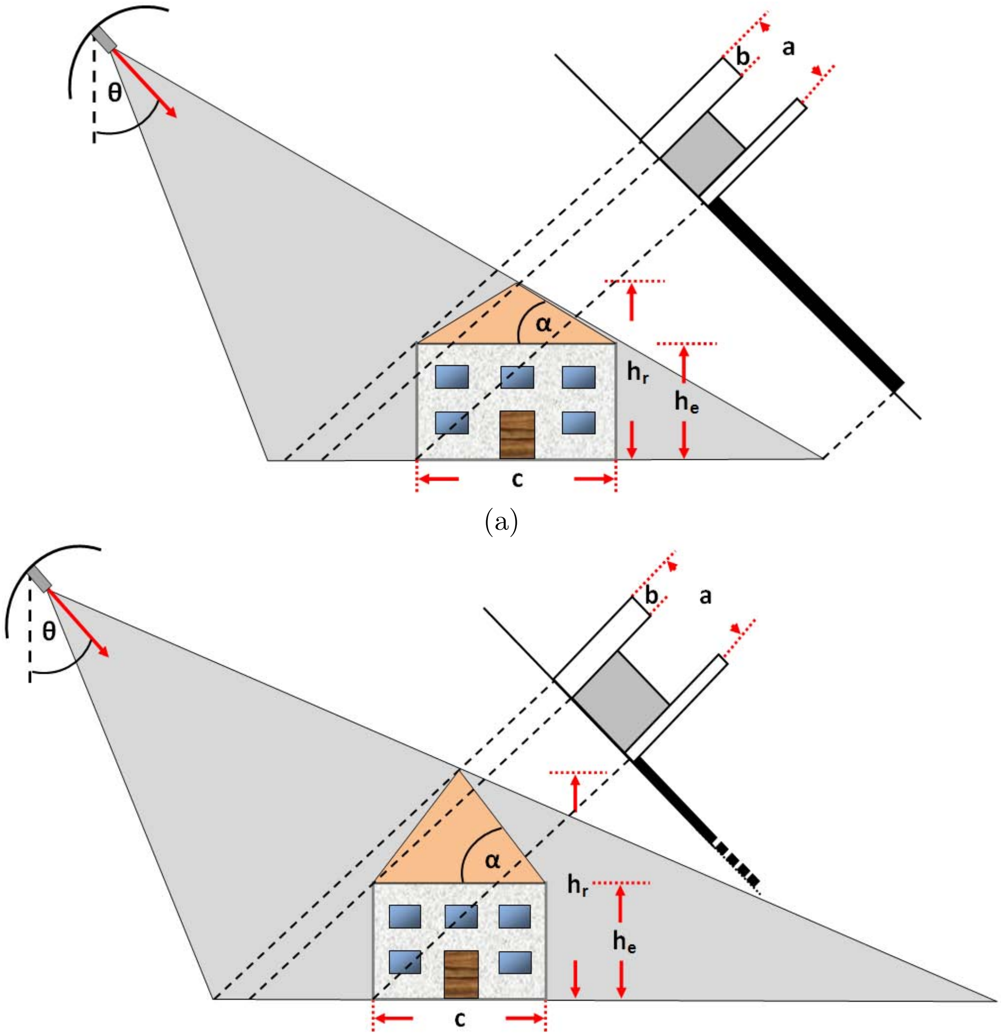


Figure 3.12.: (a) Optical image of a gable roof building, the eave overlaps with the SAR double-bounce line (red), (b) corresponding SAR image where the double-bounce line is the bottom (far range) white line, the top (near range) line is caused by direct signal reflection at the tilted roof plane directed towards the SAR sensor (range top to bottom).

measurement combining SAR double-bounce line and building roof edge in the optical image is proposed in equation 3.8. This concept only works if the building has vertical walls and a flat roof. In case of a gable roof, the building top is the roof ridge. Due to the near nadir perspective of the optical sensor, it does not overlap with the double-bounce line (Fig. 3.12(a)). Instead, the eave of the gable roof usually overlaps. Applying equation 3.8 delivers the eave height as a result instead of the ridge height. It should be noted that a gable roof eave is usually not located in the same plane as the building wall, but slightly juts out orthogonally by approximately half a meter. This potentially induced error is very small in relation to the height accuracy we may achieve and consequently the eave is assumed to be in the same plane as the building wall.

Thiele et al. [2007b, 2010a] show that gable roof buildings lead to a second bright line in SAR data, in addition to the double-bounce line, if viewed almost orthogonally by the SAR sensor (cf. section 2.1.4, Fig.2.4(g,h)). They propose to estimate building heights based on the distance between those two parallel lines from InSAR data of two orthogonal aspects. A second aspect is needed in order to determine the building width. They perform a simulation of different phase distributions according to first building hypotheses in order to solve for two alternative roof inclinations and eave heights (Fig. 3.13). Concepts presented here are based on their findings. First, their approach is briefly reviewed. Then, it is shown how we can circumvent the need for a second aspect and phase simulations by combining single aspect InSAR data with an optical orthophoto.

InSAR data of a second aspect is unnecessary because building widths can directly be measured in the orthophoto. Phase simulation is needed in [Thiele et al., 2007b, 2010a] to resolve an ambiguity occurring because two different types of gable roof buildings lead to the same pattern in the SAR data. In figure 3.13, two gable roof buildings are schematically shown leading to equal signal return in the SAR image although they are shaped differently. They differ in eave height h_e and roof inclination α. In figure 3.13(a), the near range end of the single-bounce return of the roof plane corresponds to the eave, whereas it results from the roof ridge in figure 3.13(b). It has to be decided whether α is smaller than viewing angle θ or not. To keep equations consistent, the same notation as introduced by Thiele et al. [2007b, 2010a] is used:



(a)

(b)

Figure 3.13.: Height measurements of buildings with gable roofs: (a) roof planes with small inclination ($\alpha < \theta$), (b) roof planes with high inclination ($\alpha > \theta$), the dotted radar shadow in slant range geometry signifies that unlike in (a) not the complete shadow area is shown for visualization reasons

$$\alpha < \theta, \quad h_e = \frac{a-b}{\cos\theta}, \quad h_r = h_e + \frac{c}{2}\cdot\tan\alpha, \quad \tan\alpha = \tan\theta + \frac{2b}{c\cdot\cos\theta}, \tag{3.11}$$

$$\alpha > \theta, \quad h_e = \frac{a}{\cos\theta}, \quad h_r = h_e + \frac{c}{2}\cdot\tan\alpha, \quad \tan\alpha = \tan\theta - \frac{2b}{c\cdot\cos\theta}. \tag{3.12}$$

It has to be decided whether a building has a low eave height and a steep roof or high eave and less inclined roof. This ambiguity can be solved by estimating the eave height of a building directly via shadow measurements in an optical image (Eq. 3.7). Another possibility is to exploit the overlap of eave and double-bounce line (Fig. 3.12(a)) applying equation 3.8. A fixed eave height leaves only room for one possibility, either $\alpha < \theta$ or $\alpha > \theta$ thus circumventing phase simulations.

3.2.2. Adjustment

This section deals with the stochastic adjustment framework used for estimating a single height h_b for each building. Several height measures for buildings have been introduced in the previous section, each leading to another height value. Small discrepancies naturally occur calling for a weighted adjustment according to the accuracy of each measurement. If combining different height measurements to obtain one final height value, a more accurate measurement shall gain high weighting whereas less accurate measurements shall have less influence. Accuracies of each of the previously derived building heights depend on accuracies of the corresponding observations. Double-bounce line position $\left[\begin{smallmatrix} db_x \\ db_y \end{smallmatrix}\right]$ is one observation in equation 3.8, for example. Its positional accuracy expresses how accurately db_x and db_x can be measured, depending on small errors concerning double-bounce line extraction [Wegner et al., 2010] and projection to the optical geometry. More accurate observations lead to higher influence of the entire height equation on the final height h_b. A standard approach allowing to consider observation accuracies is *least squares adjustment with functionally dependent parameters*[7], the general case of least squares adjustment. It allows to include multiple observations and unknowns within one functional relationship. Observations are, for example, distance measurements, unknowns are building heights, and functional relationships are the height equations introduced in the previous section. Much literature deals with this topic (e.g., [Mikhail, 1976, chap. 9]), only basic formulas of Niemeier [2002] will be given here, model details can be found ibidem.

The general formulation of the functional model is $\mathbf{F}\left(\hat{\boldsymbol{l}}, \hat{\boldsymbol{x}}\right) = 0$, where $\hat{\boldsymbol{x}}$ contains estimations of building height corrections and $\hat{\boldsymbol{l}}$ estimated observations after adjustment. All building height equations have to be rearranged in order to express the functional relationships implicitly. It is done by simply subtracting the building height of both sides of each equation of the previous section. Then, the left sides become zero and $-h$ appears on the right sides. The adjustment framework needs an initial building height approximation h_0 per building, which is set to the height of the most accurate measurement. It returns a correction Δh for h_0 leading to adjusted building height $\hat{h} = h_0 + \Delta h$. In fact, Δh is determined during adjustment, adjusted corrections of all buildings are contained in $\hat{\boldsymbol{x}}$ (cf. Eq. 3.18). A linearization of the functional relationships is done with a first order Taylor series expansion in parameter space. The linearized functional model of least squares adjustment with functionally dependent parameters is:

$$\boldsymbol{Bv} + \boldsymbol{A}\hat{\boldsymbol{x}} + \boldsymbol{w} = \boldsymbol{0}, \tag{3.13}$$

where $\mathbf{B}$ contains the first partial derivatives of the functional relationships with respect to the observations (Jacobian of observations) and $\mathbf{A}$ those with respect to the estimated parameters (Jacobian of heights in this case). $\mathbf{B}$ is a vector here[8] because all first partial derivatives with

[7] Least squares adjustment with functionally dependent parameters is known as *least squares adjustment with conditions* and *Gauß-Helmert-Model*, too.

[8] Considering the rearranged functional relationship of h_{db}, for example,

respect to the measured heights h are -1. Residuals $\mathbf{v} = \hat{\boldsymbol{l}} - \boldsymbol{l}$ are differences between adjusted observations $\hat{\boldsymbol{l}}$ and the initial observations $\boldsymbol{l}$. The vector of inconsistencies $\mathbf{w}$ stores differences between single height measurements h depending on observations $\mathbf{l}$ and the inital approximation h_0. The general concept of least squares adjustment is to minimize the weighted sum of squared residuals v of observations $\mathbf{l}$ with weights $\mathbf{l}$ contained in matrix $\mathbf{P}$:

$$\boldsymbol{v}^T \boldsymbol{P} \boldsymbol{v} \rightarrow min. \tag{3.14}$$

Matrix $\mathbf{P}$ is the inverse of variance-covariance matrix $\boldsymbol{Q}_l$ of observations $\mathbf{l}$. High variances in $\boldsymbol{Q}_l$ lead to low weights in $\mathbf{P}$ resulting in less influence of corresponding heights on the final adjusted value. Covariances occur if interdependencies cross-correlate observations of the functional relationships (i.e., the different height measurement equations of the previous section). No significant cross-correlations occur concerning equations 3.7 to 3.12, therefore only elements of the principal diagonal of $\boldsymbol{Q}_l$ are filled with observation variances, all other elements are zero. Those variances are exactly the accuracies of observations needed, their choice for each observation will be explained in section 4.3. Weight matrix $\mathbf{P}$ can simply be computed as $\mathbf{P} = \sigma_0 \boldsymbol{Q}_l^{-1}$, where σ_0 acts as a scaling factor. In addition to $\boldsymbol{v}^T \boldsymbol{P} \boldsymbol{v}$, the linearized functional relationships (Eq. 3.13) have to be included as conditions in the objective function. A standard technique to solve a minimization task with conditions is to use Lagrange multipliers, which are expressed indirectly via so-called correlates contained in $\mathbf{k}$. These correlates, multiplied with the linearized functional model, are added to the sum of squares resulting in the objective function to be minimized:

$$\Omega = \boldsymbol{v}^T \boldsymbol{P} \boldsymbol{v} + 2\boldsymbol{k}^T \left(\boldsymbol{B}\boldsymbol{v} + \boldsymbol{A}\hat{\boldsymbol{x}} + \boldsymbol{w} \right). \tag{3.15}$$

In order to determine the minimum, the first partial derivatives of Ω with respect to the variables $\boldsymbol{v}$ ($\boldsymbol{x}$ set to zero) and with respect to $\boldsymbol{x}$ have to be computed:

$$\frac{\partial \Omega}{\partial \boldsymbol{v}} = 2\boldsymbol{P}\boldsymbol{v} - 2\boldsymbol{B}^T\boldsymbol{k} = 0 \ , \quad \frac{\partial \Omega}{\partial \boldsymbol{x}} = -2\boldsymbol{A}^T\boldsymbol{k} = 0 \tag{3.16}$$

After rearranging and substitutions (details in [Niemeier, 2002, p. 157]) so-called normal equations are obtained:

$$\begin{bmatrix} \boldsymbol{B}\boldsymbol{Q}_{ll}\boldsymbol{B}^T & \boldsymbol{A} \\ \boldsymbol{A}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{k} \\ \hat{\boldsymbol{x}} \end{bmatrix} = \begin{bmatrix} -\boldsymbol{w} \\ \mathbf{0} \end{bmatrix} \tag{3.17}$$

This linear system has to be solved for building height corrections in $\hat{\boldsymbol{x}}$. *LU decomposition* with partial pivoting and equilibration is used here, being a direct solver based on the Gaussian elimination principle (details in [Ziehn, 2010, p. 54]). Having solved for corrections $\hat{\boldsymbol{x}}$ (elements Δh), initial

$0 = H \cdot \left(1 - \sqrt{\frac{(db_x-\nu_x)^2+(db_y-\nu_y)^2}{(e_x-\nu_x)^2+(e_y-\nu_y)^2}}\right) - h_{db}$ the first partial derivative is $\frac{\partial F(1,x_0)}{\partial h_{db}} = -1$.

building heights $\boldsymbol{x}_0$ (elements h_0) are updated to get the adjusted heights in $\hat{\boldsymbol{x}}_{\boldsymbol{a}}$ (elements correspond to $\hat{h}$, which is considered the final building height h_b).

$$\hat{\boldsymbol{x}}_{\boldsymbol{a}} = \boldsymbol{x}_0 + \hat{\boldsymbol{x}} \tag{3.18}$$

The posterior variance factor $\hat{\sigma}_b^2$ of the adjusted building height is determined following the law of error propagation:

$$\hat{\sigma}_b^2 = \frac{\boldsymbol{v}^T \boldsymbol{P} \boldsymbol{v}}{f - u}, \tag{3.19}$$

where f is the number of height measurements per building. Scalar u represents the number of unknowns, one building height per building in this case (i.e., the denominator always is $f - 1$). Taking the square root of $\hat{\sigma}_b^2$ delivers the posterior standard deviation $\hat{\sigma}_b$ of adjusted final height h_b of a building. It indicates the interior accuracy of the model, commonly known as precision. Absolute offsets to ground truth cannot be recognized directly through the adjustment framework, but have to be computed by comparing adjusted heights to reference heights. It will be done in section 4.3 by comparison to building heights acquired with airborne laserscanning thus delivering a real accuracy.

4. Experiments

In this chapter experimental results are presented and described in order to evaluate the impact of novel methods introduced in Chapter 3. First, developed contextual techniques based on Conditional Random Fields are tested (4.2) and, second, building heights are estimated (4.3). All results presented in this chapter will be discussed in Chapter 5.

Tests are conducted to assess general benefits and limitations of Conditional Random Fields compared to a standard Maximum Likelihood classifier (ML) and Markov Random Fields (4.2.1). In section 4.2.2 the impact of an irregular graph structure based on image regions is compared to a standard grid-graph on image patches. Moreover, the region-graph is exploited to test the gradient discontinuity constraint. Then, it is analysed how SAR double-bounce lines contribute to building detection (4.2.3). Furthermore, the novel ISC-CRF is applied to building detection (4.2.4) and to various images taken from computer vision benchmark datasets in order to assess whether the goal of transferability is met (4.2.5).

After object detection experiments, building heights are estimated (4.3) and compared to ground truth. The choice of standard deviations, necessary for weighting observations according to their accuracy in the least squares adjustment framework, is explained in section 4.3.1. All single measurements, their characteristics and performance, are compared in section 4.3.2. These heights are then jointly adjusted at each building using least squares adjustment with functionally dependent parameters 4.3.3.

4.1. Data

Data used for testing the proposed novel approaches is presented in this section. Building detection and height estimation are evaluated with an optical orthophoto and one mono-aspect InSAR image pair of the city Dorsten, Germany. The orthophoto was originally taken with an analogue aerial camera Zeiss RMK and scanned (©Geoinformation NRW). Pixel size on ground is 0.31 m. Single-pass X-band InSAR data (wavelength $\lambda = 3.14$ cm, baseline 2.4 m) were acquired by the AeS-1 sensor of Intermap Technologies, a description of sensor details is provided by Schwäbisch & Moreira [1999]. Spatial data resolution of the original single-look data is 0.385 meters in range and 0.18 meters in azimuth. All figures showing cut-outs of SAR images in this thesis have been extracted from the image in figure 4.2 if not stated otherwise. Optical remote sensing images are parts of the orthophoto depicted in figure 4.1.

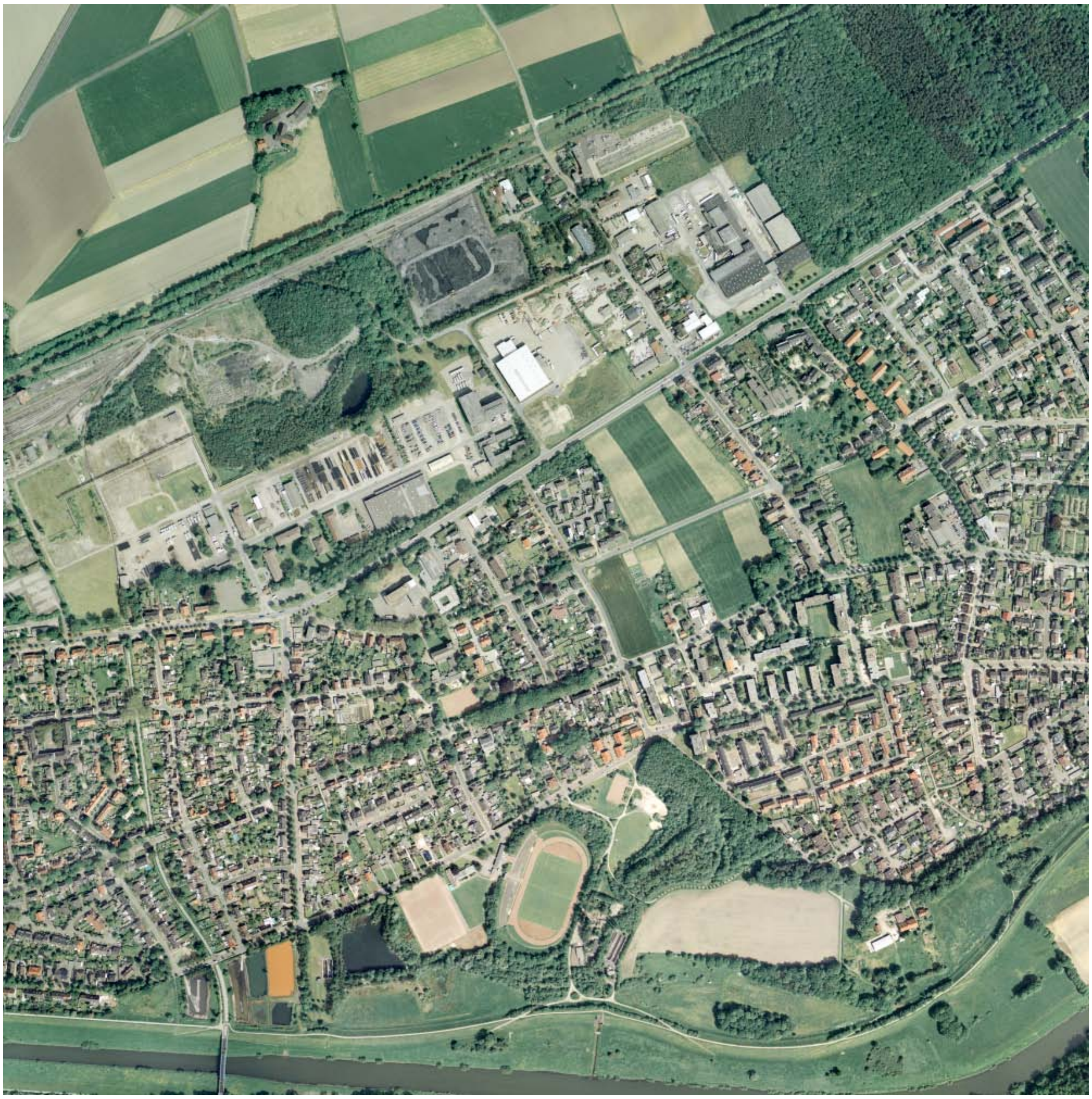


Figure 4.1.: Optical orthophoto of the Dorsten test site (©Geoinformation NRW).

For cross-validation purposes the test scene is subdivided into non-overlapping parts of 1000 × 1000 pixels (corresponding to 310 m × 310 m on ground). Performance of novel methods has to consider different building types. Large multi-storey buildings as well as single family houses with either flat roofs or gable roofs should be contained in subscenes. Those building types are mapped differently, both, in terms of their local appearance and surrounding context. Multi-storey buildings usually have greyish flat roofs, large sun shadow areas occur (caused by great building height), and long bright SAR double-bounce lines exist. Single family houses occur as different patterns. SAR double-bounce lines are not as bright because the microwave signal aggregates over a much smaller area. Furthermore, family houses usually have gable roofs leading to an additional bright line caused by single bounce at the tilted roof plane facing the SAR sensor (cf. section 2.1.4). They

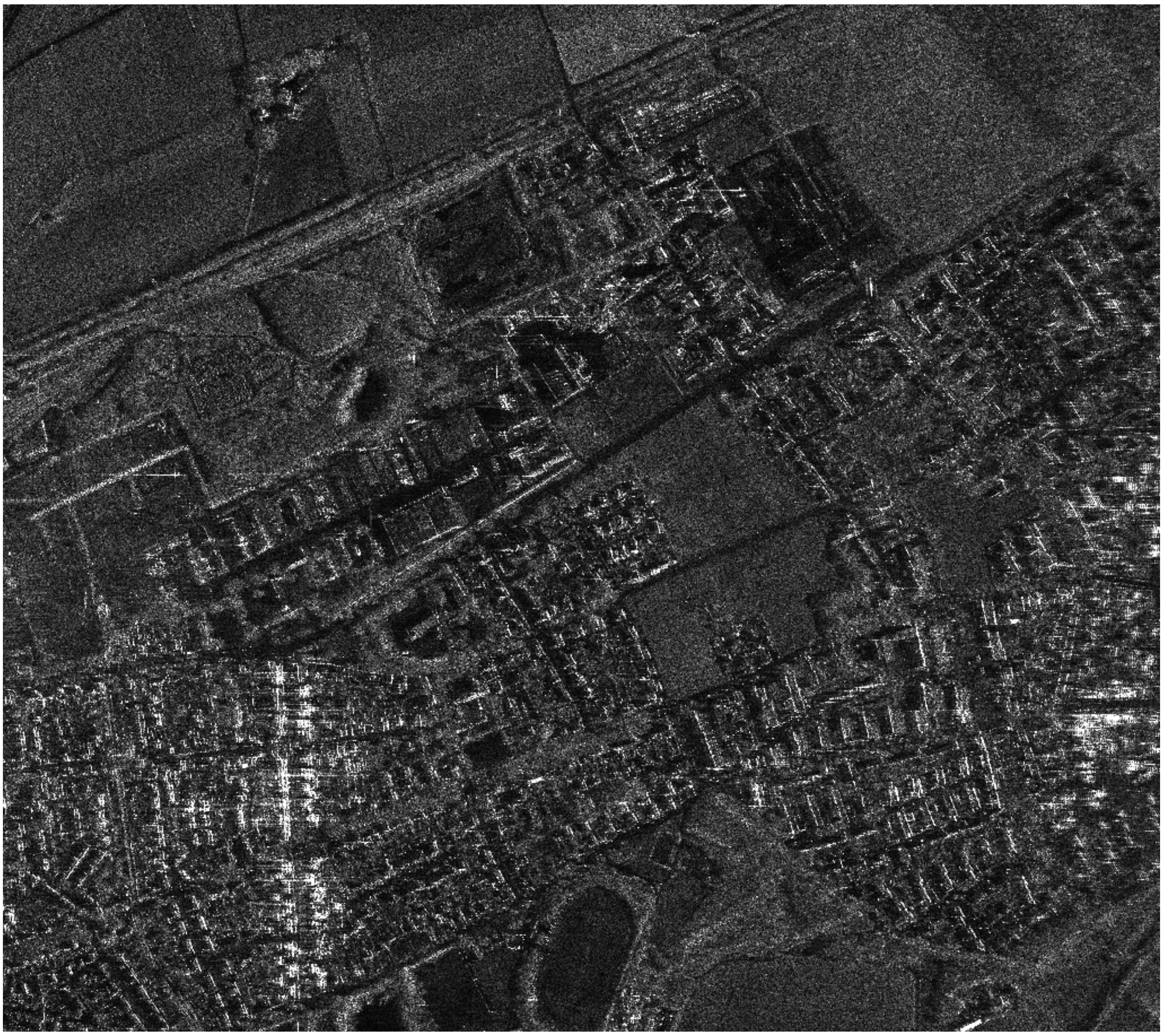
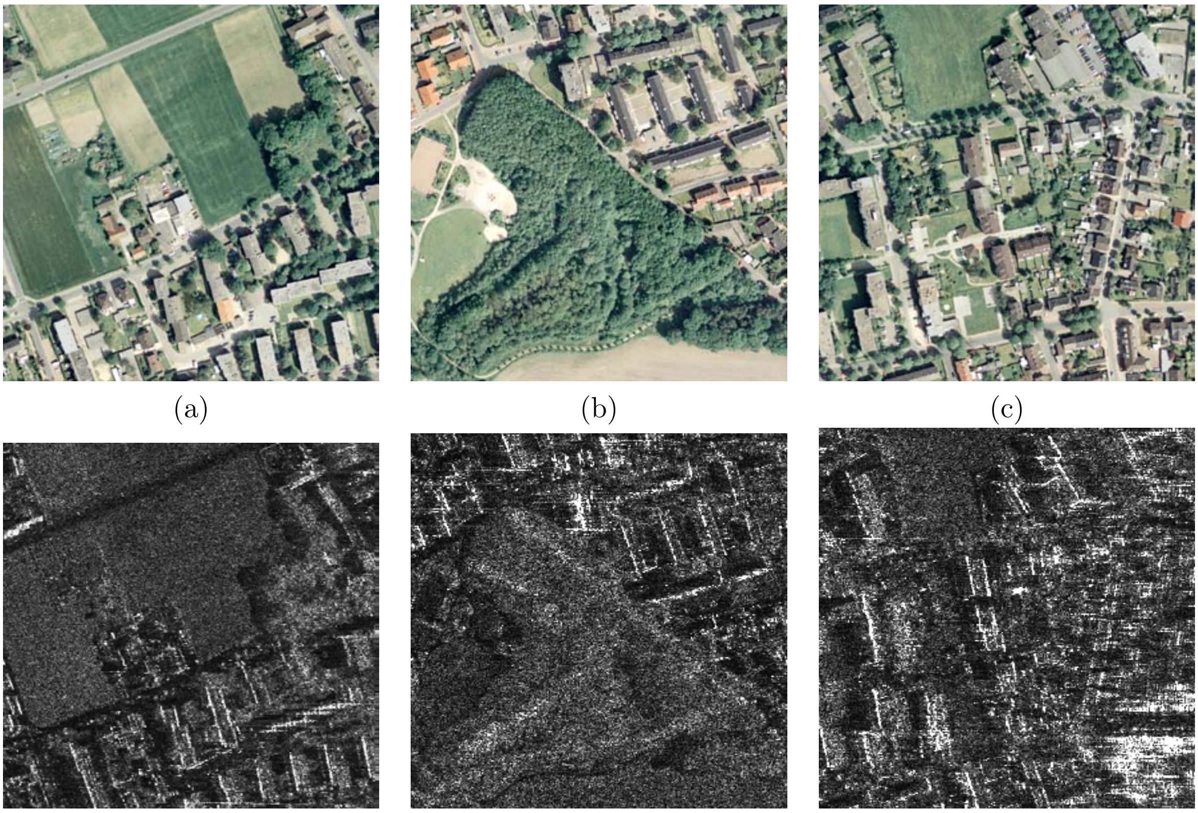


Figure 4.2.: One SAR amplitude image of the interferometric SAR image pair acquired with Intermaps AeS sensor (range direction from right to left), the image has been roughly rectified only for visualization in order to ease comparison with the optical orthophoto.

are embedded into another context compared to multi-storey buildings, too. Most small houses have front yards, driveways, gardens, and sun shadows are short, for example. Gaps between neighboring houses are narrow leading to relatively dense clusters of small buildings. Multi-storey buildings are sparsely distributed, large parking lots are located closely. Therefore, the three subscenes shown in figure 4.3 containing all building types, building distributions, and building contexts are selected for experiments. Building height estimation is conducted for different building types, too. Heights of small and medium size gable roof builings as well as big flat roof buildings are estimated based on test data shown in figures 4.1 & 4.2. All processing is done on a computer with Intel$^{\mathrm{TM}}$ Core i7 2.4 GHz CPU and 12 GB RAM.



(a) (b) (c)

(d) (e) (f)

Figure 4.3.: (a)-(c) Subscenes of the optical orthophoto (Fig. 4.1) and (d)-(f) corresponding subscenes of one SAR magnitude image (Fig. 4.2) of the interferometric pair.

4.2. Object detection results

Object detection results of the developed approaches based on CRFs are presented in this section. Buildings are detected combining optical and SAR features, tests with a simulated urban scene are conducted, and the ISC-CRF is applied to images of computer vision benchmark data. For all object detection experiments limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) [Nocedal, 1980; Liu & Nocedal, 1989; Nocedal & Wright, 2006], a quasi-Newton method, is applied for parameter estimation and loopy belief propagation (LBP) [Frey & MacKay, 1998] for approximate inference[1]. All implementations are based on a MatLab/C-toolbox called CRF2D [Vishwanathan et al., 2006b] for binary classification tasks with CRFs originally used in Vishwanathan et al. [2006a].

Features generated from optical images

All building detection experiments are conducted with the three subregions of figure 4.3. A first impression of suitable features is achieved comparing marginal distributions of each single feature of classes building and background. Features showing very distinct marginals for both classes are

[1] Details of parameter estimation and inference are described in Annex A.

taken as basic features. Difference of marginals is a rather weak feature selection technique because it does not take into account joint distributions. Therefore, basic features are combined with various other features that did not perform well regarding the difference of marginals. Lots of configurations are tested within the CRF framework. According to this rather experimental and simple feature selection technique,

- mean of red and green channel (normalized by the length of the RGB vector),
- hue mean and standard deviation,
- and saturation mean

are found to be descriptive colour features. Additional features are generated based on gradient orientation histograms of the intensity image [Dalal & Triggs, 2005] as already used for detection of building facades [Kumar & Hebert, 2006; Korč & Förstner, 2008]. Slightly different features are derived because it is dealt with aerial imagery and no facades with characteristic horizontal and vertical gradients appear. Second and third central moments (variance and skewness) of gradient orientation histograms are used as features. Each feature is computed within the spatial unit (i.e., square image patch or image region generated via segmentation) that is represented by a single node in the graph (cf. Fig. 2.8).

Building features in InSAR data

Double-bounce lines extracted in InSAR data (cf. section 2.1.4, Fig. 4.4(a)-(c)) or a single SAR image [Wegner et al., 2010] are building hints in SAR data. In this thesis, lines segmented in InSAR data are exploited for testing, but lines of single SAR images may be used, too[2]. Rule-based approaches directly detect and reconstruct buildings based on extracted points, lines, or polygons [Stilla, 1995; Michaelsen & Stilla, 2002; Soergel et al., 2003b], but object detection viewed from a classification perspective needs scalar values to be written into feature vectors.

In order to turn hints into features, distance maps are generated from double-bounce lines in range direction of the SAR sensor (Fig. 4.4(d)-(f)). Lines occur directly where building walls facing the SAR sensor meet ground. It is very likely that pixels located directly behind a line in range direction belong to a building (cf. Fig. 4.4(a)-(c)). Assuming a certain maximum extent of buildings in the scene, double-bounce lines cannot make any significant prediction beyond this distance. Building evidence decreases from value one to zero with increasing distance to the double-bounce line in range direction. Maximum building extent is set to 70 pixels, approximately 21.7 m on ground. This distance is chosen as a compromise between width and length of buildings because lines can occur either at the long or the short side of a building.

[2] As described in section 2.1.4, segmented double-bounce lines are projected to orthophoto geometry with InSAR heights. In case of a single SAR image, assuming one fixed height for the entire scene and flat terrain enables line projection. Another possibility would be the introduction of a DEM, for example derived from airborne laserscanning data. One could also measure corresponding points in orthophoto and SAR image, transform the SAR image to orthophoto geometry, and segment lines. It would, however, result in distorted double-bounce lines.

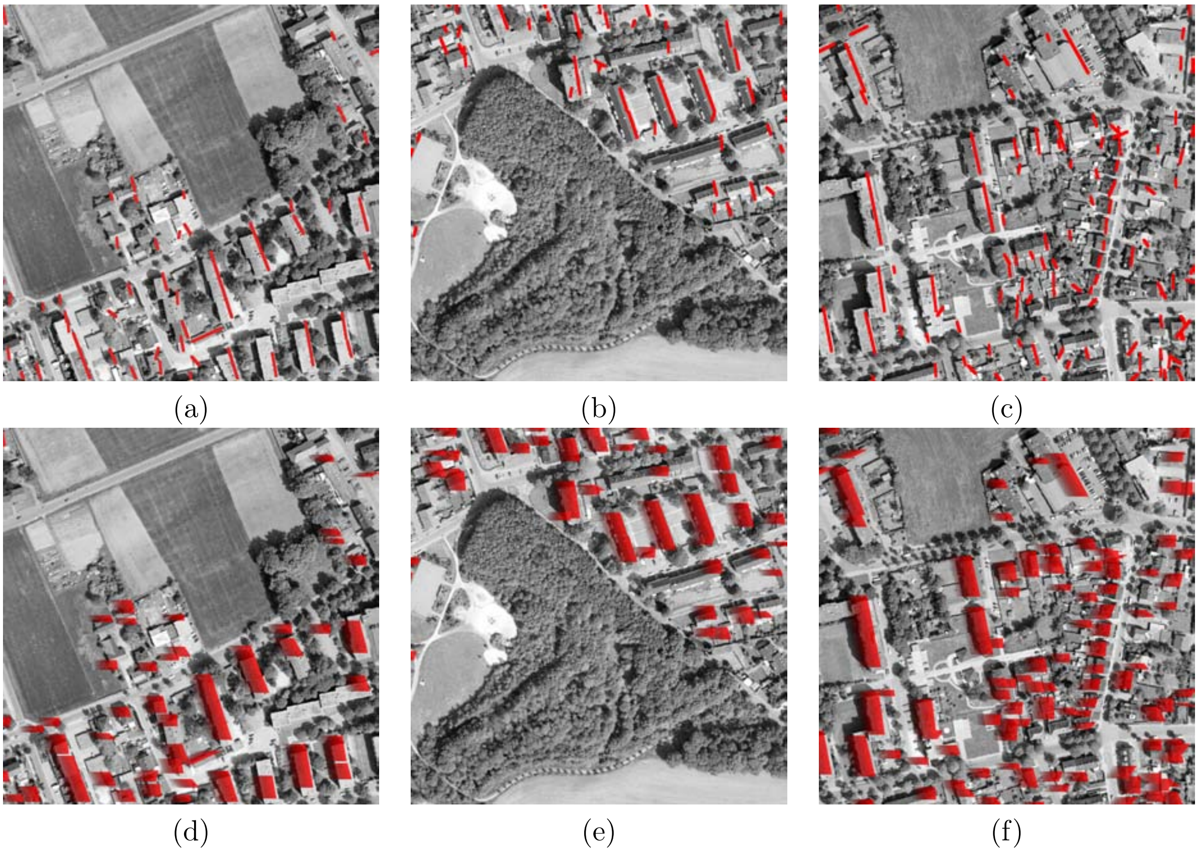


Figure 4.4.: InSAR double-bounce lines as building hints: (a)-(c) Intensity channel of optical aerial images overlaid with lines segmented in subscenes shown by one of the SAR magnitude images per interferometric pair in figures 4.3, (d)-(f) distance maps overlaid to the optical images (maximum building extent 70 pixels ≈ 21.7 m); range direction of SAR sensor right to left.

Features are derived from distance maps (shown red in Fig. 4.4(d)-(f)) by calculating scalar values for each patch or region, which can be imagined as superimposing a patch grid or segmentation onto the distance map. Five features are computed per node (i.e., within a patch or region): maximum, mean, median, standard deviation, and percentage of non-zero entries. The last feature counts the percentage of pixels of a patch or region being covered by double-bounce line evidence.

Feature vector expansion

All optical and SAR features are scaled between zero and one. A quadratic expansion of feature vectors $\mathbf{h}_i(\mathbf{x})$ is done as described by Kumar & Hebert [2006], who state that this step may be viewed as a *kernel mapping of the original feature vector into a high dimensional space.* It introduces a quadratic decision surface in feature space capable of more precisely discriminating building nodes from background nodes compared to a simple linear one. The basic idea is that a linear classifier, as used in this thesis (cf. Eq. 2.17 & 2.18), applied in a quadratically expanded feature space will yield a quadratic decision surface in original feature space. Simple linear models can be kept, allowing for efficient parameter estimation by introducing a higher order feature space. A quadratic feature vector contains all original elements, their squares, and pairwise products. Kumar & Hebert [2006] mention that this is *equivalent to the kernel mapping of the data using a polynomial kernel of degree two.* Each first component of an expanded node feature vector is set to one in order to accommodate a so-called bias parameter, which is the first element of the corresponding weight vector. Its effect can be interpreted as shifting the decision surface in feature space, exact shape modelling is done by all other parameters.

Evaluation strategy

Results of all classification experiments are evaluated in terms of *false positive rate* (FPR) and *true positive rate* (TPR). FPR is the percentage of all background pixels being misclassified as building pixels. TPR represents the percentage of all building pixels being correctly classified as such. In general, the goal is to develop a classification technique delivering results with high TPR and low FPR. In order to ease visual interpretability of results, CRF classification outcomes are overlaid to the intensity channel of the optical image. False positive pixels are coloured red, true positive pixels green, missed building pixels blue (false negatives), and correctly classified background (true negatives) without any colour (cf. Fig. 4.5).

Cross-validation is performed for all experiments in order to avoid particular training/testing-setups biasing classification results. Corresponding to Crowther & Cox [2005], the optimum experimental setup is to use two thirds of data for training and one third for testing. Thus, three-fold cross-validation is conducted, two subscenes for training and one subscene for testing (considering subscenes of figure 4.3). Each experiment is done three times with changing training and testing image combinations.

Dealing with the three subscenes chosen for building detection experiments (Fig. 4.3), the first processing run takes features of 4.3(a,d) & (b,e) for training and those of 4.3(c,f) for testing. Classification results of figures 4.3(c,f) in terms of TPR and FPR are recorded. A second processing run (i.e., second fold) is conducted with features of 4.3(b,e) & (c,f) for training, 4.3(a,d) for testing. TPR and FPR of building detection results of 4.3(a,d) are recorded. For the third fold of cross-validation, the final combination with 4.3(a,d) & (c,f) for training and 4.3(b,e) for testing is processed, results are recorded. The mean of all three cross-validation folds of TPR and FPR is computed and reported as final result.

4.2.1. CRF versus Maximum Likelihood and Markov Random Field

In order to assess quality and characteristics of the CRF method, building detection results are compared to two other probabilistic supervised learning methods: Maximum Likelihood (ML) and Markov Random Fields (MRF). ML is a generative standard approach for classification based on the Bayesian theorem. The main difference of ML compared to MRF is that it does not model local context information through the prior term (cf. section 2.2.3). ML simply assumes a uniformly distributed prior and maximizes the likelihood term. A multivariate Gaussian function is used, the standard approach, to model the likelihood. It may be seen as a baseline approach of probabilistic classification without using prior (context) information. MRF is a state-of-the-art contextual classification technique. Context is considered through the prior term (cf. section 2.2.3).

Optical (colour and texture) and InSAR (double-bounce line distance maps) features, as described in the previous section, are input to ML, MRF, and CRF. A grid-graph based on square image patches of size 20 × 20 pixels is used. Building detection results of ML, MRF, and CRF based on a regular graph of image patches are summarized in table 4.1, results on an irregular region-graph in table 4.2. ML results are shown in figure 4.5(a)-(f), MRF results in figure 4.5(g)-(l), and CRF results in figure 4.7(a)-(f).

Classifier	TPR [%]	FPR [%]	Time [min]
Maximum Likelihood	70.7	16.1	1.5
Markov Random Field	88.4	31.4	11.5
Conditional Random Field	77.6	23.4	21.5

Table 4.1.: TPR, FPR, and computation time per image achieved on a patch graph with ML, MRF, and standard CRF as described in section 2.2.4.

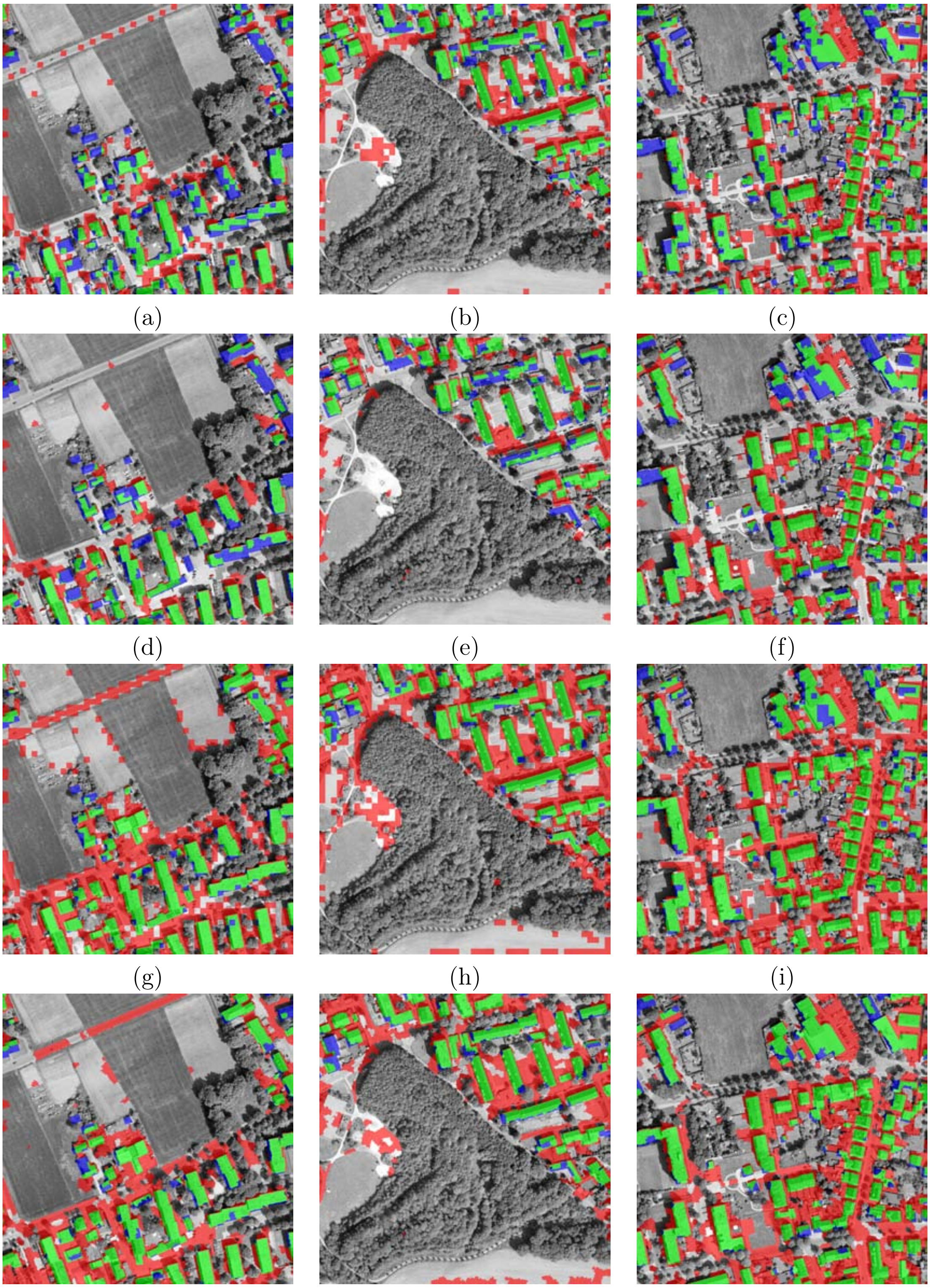
Classifier	TPR [%]	FPR [%]	Time [min]
Maximum Likelihood	68.8	14.2	1.0
Markov Random Field	86.0	26.8	2.5
Conditional Random Field	79.5	22.0	1.7

Table 4.2.: TPR, FPR, and computation time per image achieved on a region-graph with ML, MRF, and standard CRF as described in section 2.2.4.

4.2.2. Patches versus regions

In this section, CRF performance with an irregular graph structure based on image regions is tested. Various segmentation methods have been tried: a watershed segmentation of the intensity channel of the optical images, Normalized Cuts [Shi & Malik, 1997, 2000], Mean Shift [Comaniciu & Meer, 1999, 2002], and Quickshift [Vedaldi & Soatto, 2008].

A standard watershed segmentation is fast to process, but only considers gradients, resulting in the scene topology not being well captured. Normalized Cuts consider spatial proximity and colour in addition to a gradient-based constraint, but are computationally very costly in terms of processing time and memory. Segmentation tests with an improved version (with respect to the originally proposed algorithm of Shi & Malik [1997, 2000]) considering multiple scales to speed up computation [Cour et al., 2005] led to very long computation times. In order to partition each test image of size 1000×1000 pixels shown in figure 4.3 into 950 regions, needed for over-segmentation to preserve scene details, two weeks per image are necessary. Choosing smaller images accelerates computation, but to learn meaningful contextual links, images must not be smaller than 1000 × 1000 pixels (considering



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure 4.5.: Building detection results: ML on patches (a)-(c) and on regions (d)-(f); MRF on patches (g)-(i) and on regions (j)-(l); based on optical and double-bounce line features; corresponding standard CRF building detection results are shown in figures 4.7(g)-(i).

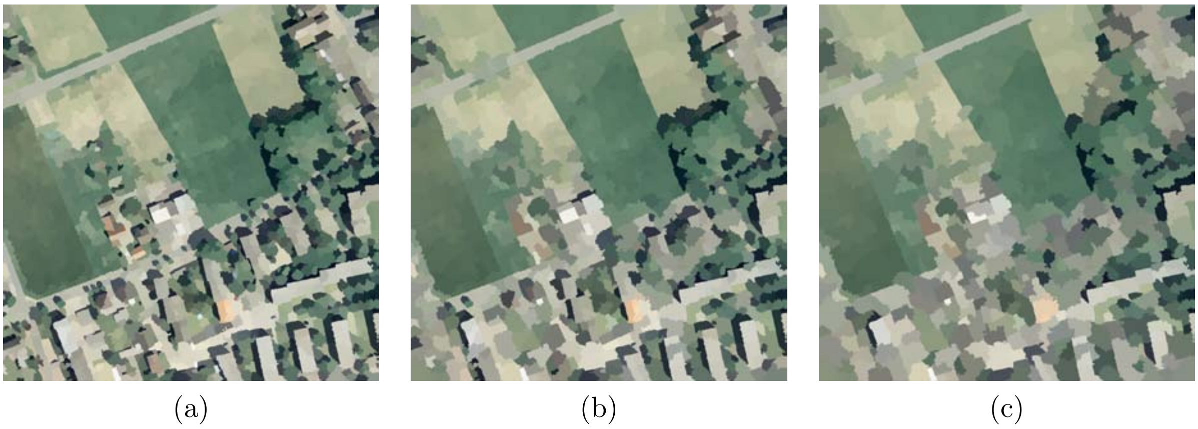


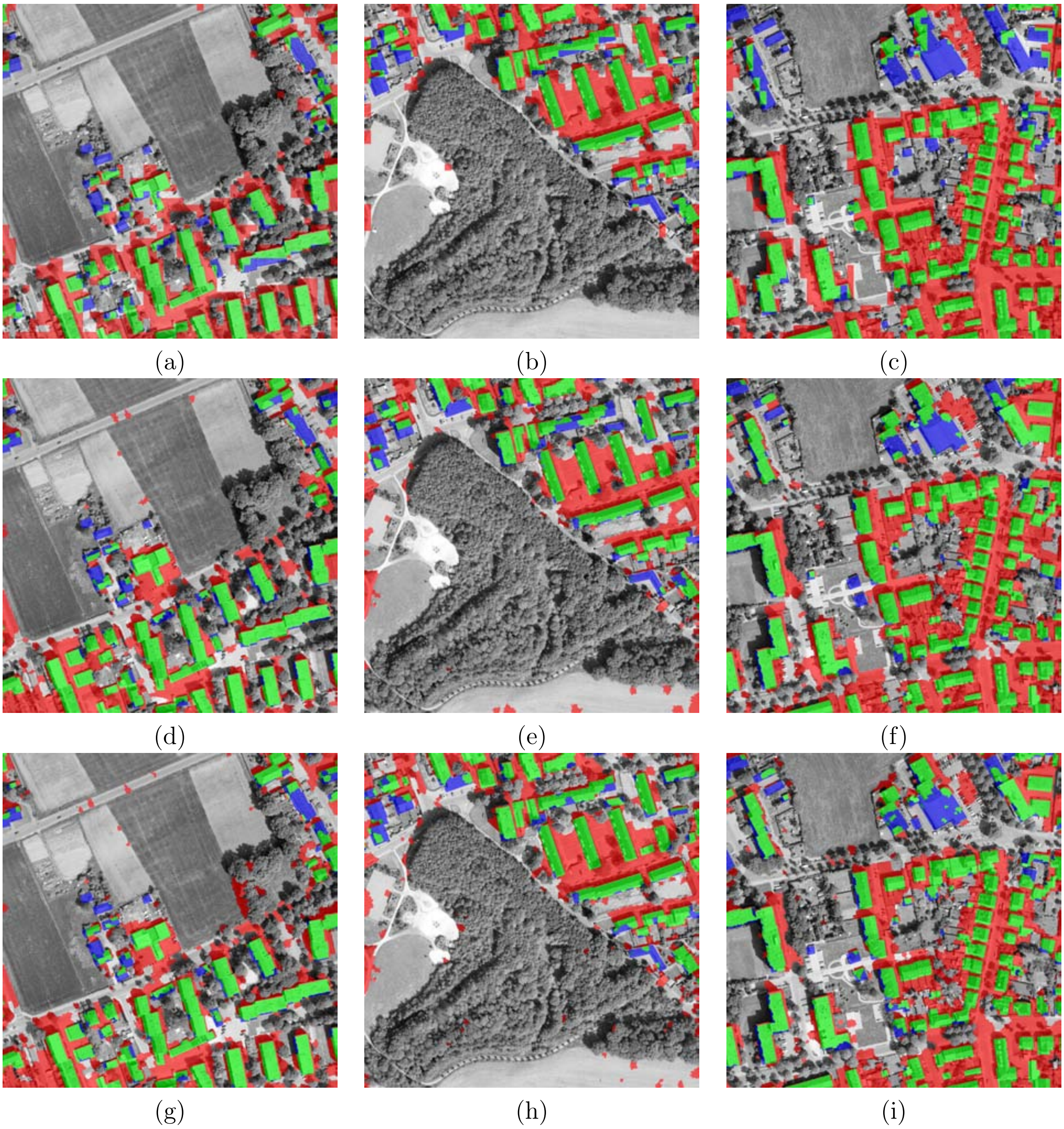
Figure 4.6.: Quickshift segmentation at three scales of the image presented in figure 4.3(a) with kernel size $ks = 3$, ratio $ra = 0.2$, and cuts through the tree at (a) $md = 10$, (b) $md = 15$, and (c) $md = 18$; its mean colour is assigned to each region.

the orthophoto of Fig. 4.1). Mean Shift considers the same features as Normalized Cuts except the gradient. It is very fast to compute (several seconds per image), but regions at multiple scales cannot be derived as easily.

Quickshift is very similar to Mean Shift, being fast to compute and aggregating pixels to regions based on a four-dimensional feature space defined by spatial distances in an image, hue, saturation, and intensity. It generates a tree-like graph structure, arranging regions in descending order in terms of their size [Vedaldi & Soatto, 2008]. Biggest regions, containing all smaller regions at finer scales, function as root of the tree, the smallest ones are leaves. Regions at arbitrary scales can be obtained by simply cutting through the tree at desired resolutions without having to process anew. Small regions at finer scales do not overlap with coarser ones (cf. section 3.1.2). Thus, Quickshift allows a straightforward multi-scale segmentation without regions of different scales overlapping. It has already been used successfully for object detection with CRFs, for example by Fulkerson et al. [2009], and for building detection (e.g., [Kluckner & Bischof, 2010]). To balance pros and cons, Quickshift is chosen for all tests.

Three Quickshift parameters have to be set [Vedaldi & Soatto, 2008]: kernel size ks, ratio ra, and maximum distance md. A filter kernel is shifted across the image to compute feature distributions. Increasing ks leads to more representative distributions at the cost of increasing computation time (and smoothing of details), too. Ratio ra adjusts the tradeoff between spatial proximity and colour features, larger values giving more importance to colour. Parameter md defines the maximum distance between two points in feature space still belonging to the same region, higher values md result in larger regions. Segmentations of the optical image shown in figure 4.3(a) at three scales are given in figure 4.6.

Only the highest scale (Fig. 4.6(a)) with parameters $ks = 3$, $ra = 0.2$, and $md = 10$ is used for single-scale region-based CRF results. All further multi-scale experiments are performed based on those three scales presented in figure 4.6 ($ks = 3$, $ra = 0.2$, $md = (10, 15, 18)$). It should be



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 4.7.: CRF building detection results: (a)-(c) with a graph of square patches size 20 × 20 pixels, (d)-(f) a graph on image regions at one scale, (g)-(i) a graph on image regions at three scales.

noted that only the highest segmentation scale (i.e., smallest regions), depicted in figure 4.6(a), is represented with nodes in the graph. Features of lower scale (larger regions) are concatenated with the feature vectors of smaller regions they contain. Building detection results based on patch-graph (Fig. 4.7(a)-(c)), region-graph (Fig. 4.7(d)-(f)), and region-graph with multi-scale features (Fig. 4.7(g)-(i)) are summarized in table 4.3.

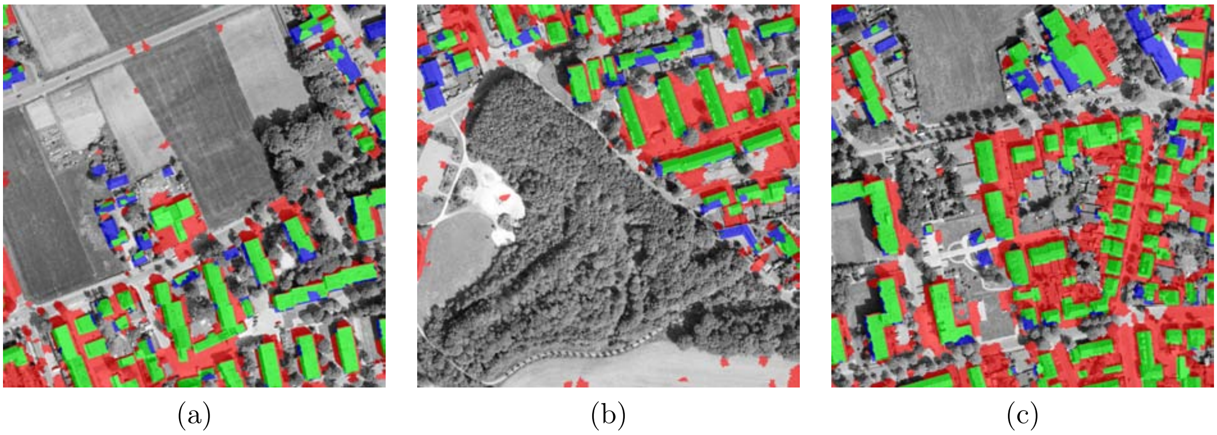


Figure 4.8.: CRF building detection results with gradient-based discontinuity constraint on region-graph.

CRF graph	TPR [%]	FPR [%]	Time [min]
Patches	77.6	23.4	21.5
Regions	79.5	22.0	1.7
Multi-scale regions	80.7	21.2	2.0

Table 4.3.: True positive rate (TPR), false positive rate (FPR) in %, and computation time per image achieved with the standard CRF as described in section 2.2.4 with a graph based on patches, regions, and regions of multiple scales.

A graph based on regions enables the introduction of a discontuity constraint as a function of the mean image gradient between two neighbouring regions (cf. section 3.1.2). 78.4 % TPR and 20.8 % FPR are achieved, results are shown in figure 4.8.

4.2.3. Impact of SAR double-bounce line

It is tested whether InSAR double-bounce lines lead to any improvement of overall building detection quality. Results of the standard CRF (cf. section 2.2.4) with and without double-bounce line features are compared. Results using only optical features are provided in figure 4.9.

InSAR double-bounce lines and generated distance maps are given in figure 4.4. Images of building detection combining SAR and optical features are presented in figures 4.7(d)-(f). TPR and FPR estimated by three-fold cross validation are shown in table 4.4.

Features	TPR [%]	FPR [%]
Optical	79.1	21.9
Optical & double-bounce line	79.5	22.0

Table 4.4.: Contribution of double-bounce lines to building detection via the standard CRF (cf. section 2.2.4) based on an irregular region-graph.

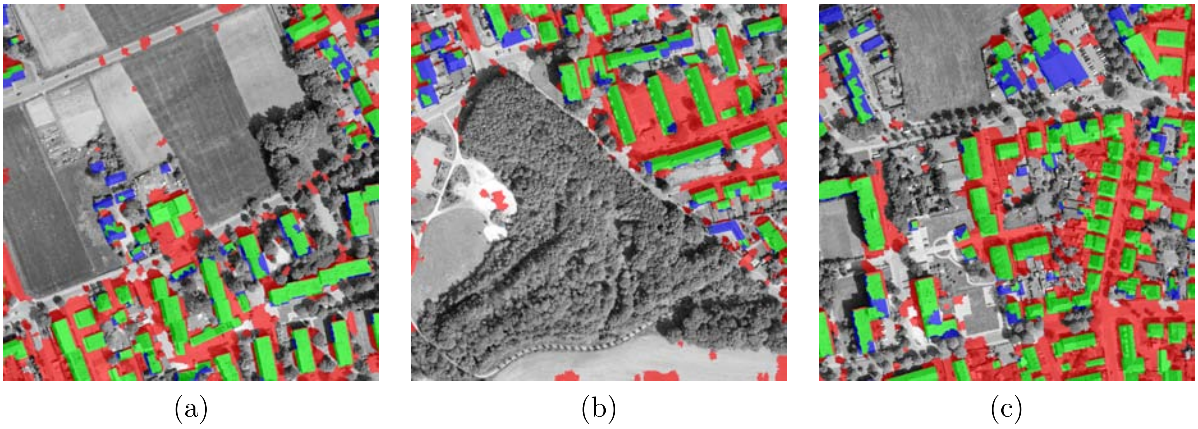


Figure 4.9.: CRF results based only on optical features; buildings detected with a CRF including double-bounce line features, too, are shown in figures 4.7(d)-(f); double-bounce lines and corresponding distance maps are provided in figure 4.4.

4.2.4. Implicit scene context

In order to assess benefits and limitations of the proposed ISC-CRF (cf. section 3.1.3), several experiments are conducted. First, standard CRF and ISC-CRF are compared using a simulated test scene where the exact number of subcategories is known a priori. Second, consequences of varying numbers of k-means cluster centers are investigated. Third, robustness to noise in comparison to the standard CRF is tested. Fourth, buildings are detected with the ISC-CRF combining optical and SAR data(Fig. 4.3).

Building detection in this thesis is viewed as a binary classification task of category building versus category non-building (i.e., background). We can interpret training data, being semantically annotated with only two classes, as partially labeled. Both classes consist of multiple subcategories (cf. section 3.1.3), which are not explicitly annotated semantically in the training database. Implicit scene context captures these subcategories via clustering and learns contextual links between different subcategories of buildings and background. Implicit patterns in data, for example, shadows next to buildings and driveways connecting single family houses with streets, can potentially be learned. Exact types and number of subcategories and their interrelations do not have to be known a priori. At each segmentation scale, the following ISC features are computed:

- Closest and second closest cluster centers to node of interest (two features),
- minimum, maximum, median, and standard deviation of occurring cluster indices at each context range (twelve features in case of three context ranges),
- most often and second most often occurring indices at each range (six features in case of three context ranges).

Twenty ISC features are computed in total at each segmentation scale if considering three context ranges. A multi-scale segmentation with three scales leads to 60 ISC features being written to a node at highest scale. ISC-CRF and standard CRF are first applied to three simulated subscenes

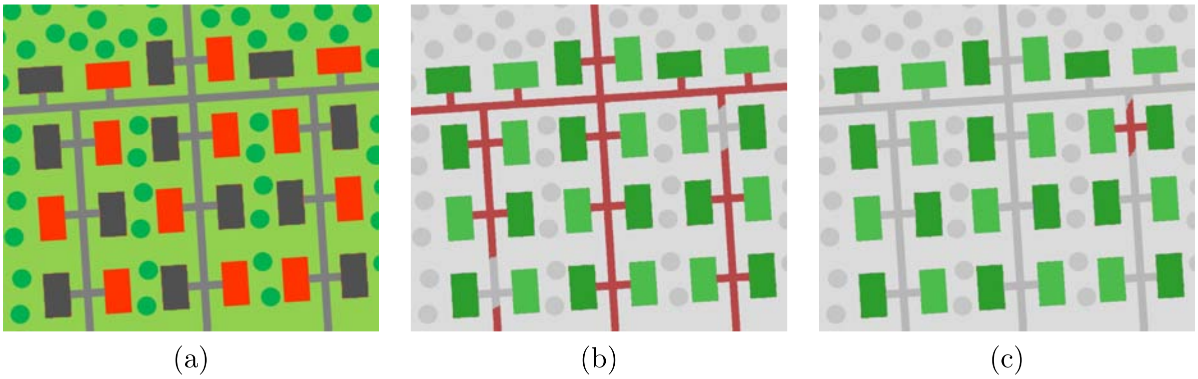


Figure 4.10.: CRF results with simulated data: (a) one image of the simulated test images, (b) detected buildings without implicit scene context (c) and with implicit scene context.

(one is shown in Fig. 4.12)(a)) containing red buildings, grey buildings, trees (dark green circles), grassland (light green background), and streets (light grey lines). Only colour features are used because no texture was simulated. Grey buildings and grey streets are closely located in feature space and thus context has to support discriminating buildings from streets. Implicit scene context is captured in three ranges (radii 10, 20, and 30 pixels) and concatenated with original colour features for ISC-CRF classification. Three-fold cross-validation is conducted and mean TPR and FPR are computed. Standard CRF (Fig. 4.12(b)) and ISC-CRF (Fig. 4.12(c)) achieve the same TPR of 85.9%. The standard CRF misclassifies 6.8% background pixels as building whereas the ISC-CRF has a significantly lower FPR of 0.8% (Tab. 4.5).

	CRF		ISC-CRF	
Data	TPR [%]	FPR [%]	TPR [%]	FPR [%]
Simulation	85.9	**6.8**	85.9	**0.8**
Optical & double-bounce line	79.5	**22.0**	78.1	**21.2**

Table 4.5.: TPR and FPR of classification results of the simulated scene (Fig. 4.12) and real remote sensing data (Fig. 4.3) achieved with standard CRF and ISC-CRF.

Cluster center number as well as segmentation scales are currently adapted manually to each data set, whereas context radii are set as a function of the mean region size of an image. The simulated urban scene (Fig. 4.12(a)) is used to evaluate the impact of varying cluster centers because the exact number of subcategories is known: red buildings, grey buildings, trees (dark green circles), grassland (light green background) and streets (light grey lines). Only colour features are used for these tests leading to five distinct clusters. Three ranges (radii 10, 20, and 30 pixels) are chosen and experiments with five up to 50 cluster centers are conducted. FPR of each ISC-CRF classification is displayed in blue Fig. 4.11(a) and such of the standard CRF in red.

Robustness of the ISC-CRF to noise is experimentally evaluated. Several Gaussian noise levels with mean zero and standard deviations up to 100 % (corresponding to 256 in our case of 8 bit RGB channels) are generated and added to RGB channels of the simulated data, which is then cropped in order to keep all values between zero and 255. Cross-validation tests with standard CRF and ISC-CRF are done and FPR is recorded. In figure 4.11(b) FPR of standard CRF (red) and of ISC-CRF (blue) considering all tested noise levels are displayed.

Finally, the ISC-CRF is applied for building detection combining optical and InSAR double-bounce line features. Implicit scene context features computed in three ranges (radii 6, 11, and 18 pixels) and ten cluster centers are set. For ISC-CRF classification they are concatenated to the original colour and texture features. TPR and FPR are given in table 4.5. ISC-CRF results are visually shown in figure 4.12, those of a standard CRF with exactly the same configuration (except ISC features) in figures 4.7(d)-(f).

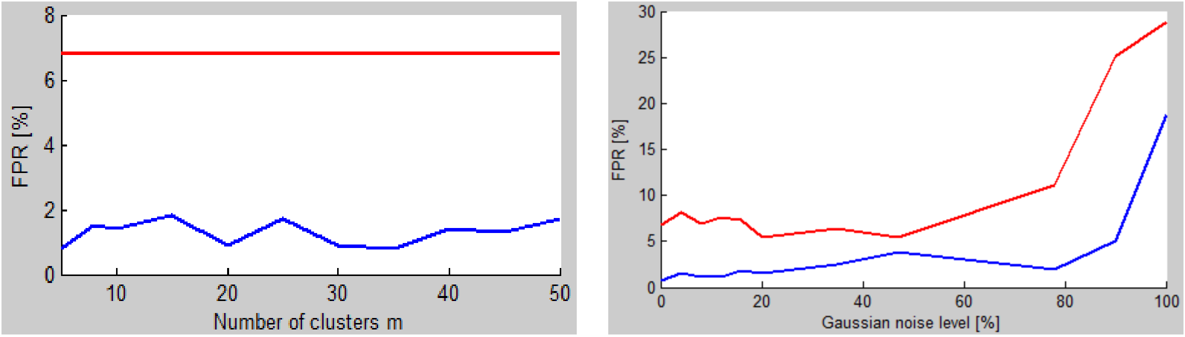
4.2.5. Transferability

In order to verify the general applicability of the implicit scene context concept to images containing any kind of objects and scene, tests are performed with three different object scenes:

- Facade images taken from the eTRIMS benchmark data [Korč & Förstner, 2009],
- images of algae found with Google$^{\mathrm{TM}}$ on the internet,
- car images of the Graz-02 benchmark data [Opelt et al., 2006].

Those particular object class categories are chosen because they represent different spatial object and background distributions. Cars are small irregular objects entirely surrounded by background context (Fig. 4.13(a)). Building facades are single very large objects with clear straight boundaries and background context only above and below (Fig. 4.13(d)). Large but frayed objects partially surrounded by background context are algae (Fig. 4.13(g)). A good performance of the implicit scene context approach for all tasks would support the claim of general applicability to different image scenes.

Experiments are conducted with nine images of each scene category, which are randomly partitioned into groups of three images for three-fold cross-validation. Example images and corresponding



(a) (b)

Figure 4.11.: FPR of ISC-CRF (blue) based on simulated data (FPR of standard CRF drawn in red): (a) with varying numbers of cluster centers and (b) with different noise levels (cluster center number fixed to five).

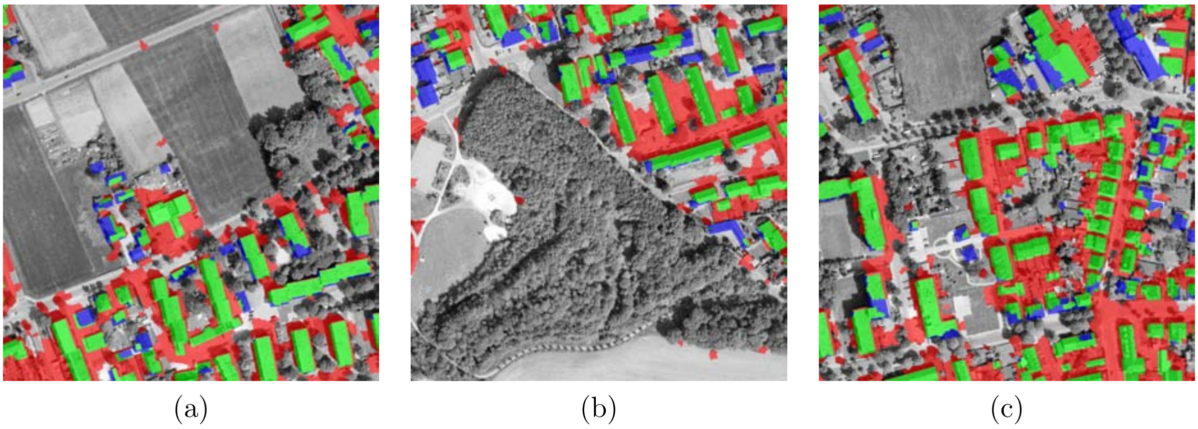


Figure 4.12.: ISC-CRF results of orthophoto and double-bounce line features (three ranges, ten cluster centers) based on a single-scale segmentation into regions; corresponding standard CRF results are given in figures 4.7(d)-(f).

	CRF		ISC-CRF	
Data	TPR [%]	FPR [%]	TPR [%]	FPR [%]
eTRIMS facades	86.9	**22.1**	88.1	**7.3**
Algae	75.7	**37.0**	84.5	**23.7**
Graz-02 cars	86.6	**16.4**	88.1	**4.3**

Table 4.6.: TPR and FPR for different objects and context patterns achieved with a standard CRF and with an ISC-CRF.

results are shown in figure 4.13. Classification performance is summarized in table 4.6. Furthermore, tests with these computer vision scenes facilitate judging to what degree of complexity context can be learned with the ISC-CRF. Real remote sensing data of an urban scene has a very high context complexity, scenes tested here have a lower complexity, and the simulated scene represents the lowest context complexity level of the ones tested.

4.3. Building height estimation results

Results of building height estimation are presented in this section. First, observation accuracies, needed for least squares adjustment with functionally dependent parameters, are derived and described. All height measurement possibilities, introduced in section 3.2.1, are tested and results are presented in section 4.3.2. Adjusted heights considering all available single measurements per building and observation accuracies are provided in section 4.3.3.

4.3.1. Accuracies of observations

In order to weight all height measurements according to their accuracy, standard deviations σ have to be assigned to each observation (cf. section 3.2.2) considered in functional relationships (cf. section 3.2.1). They will be entered as elements to the principle diagonal of variance-covariance matrix

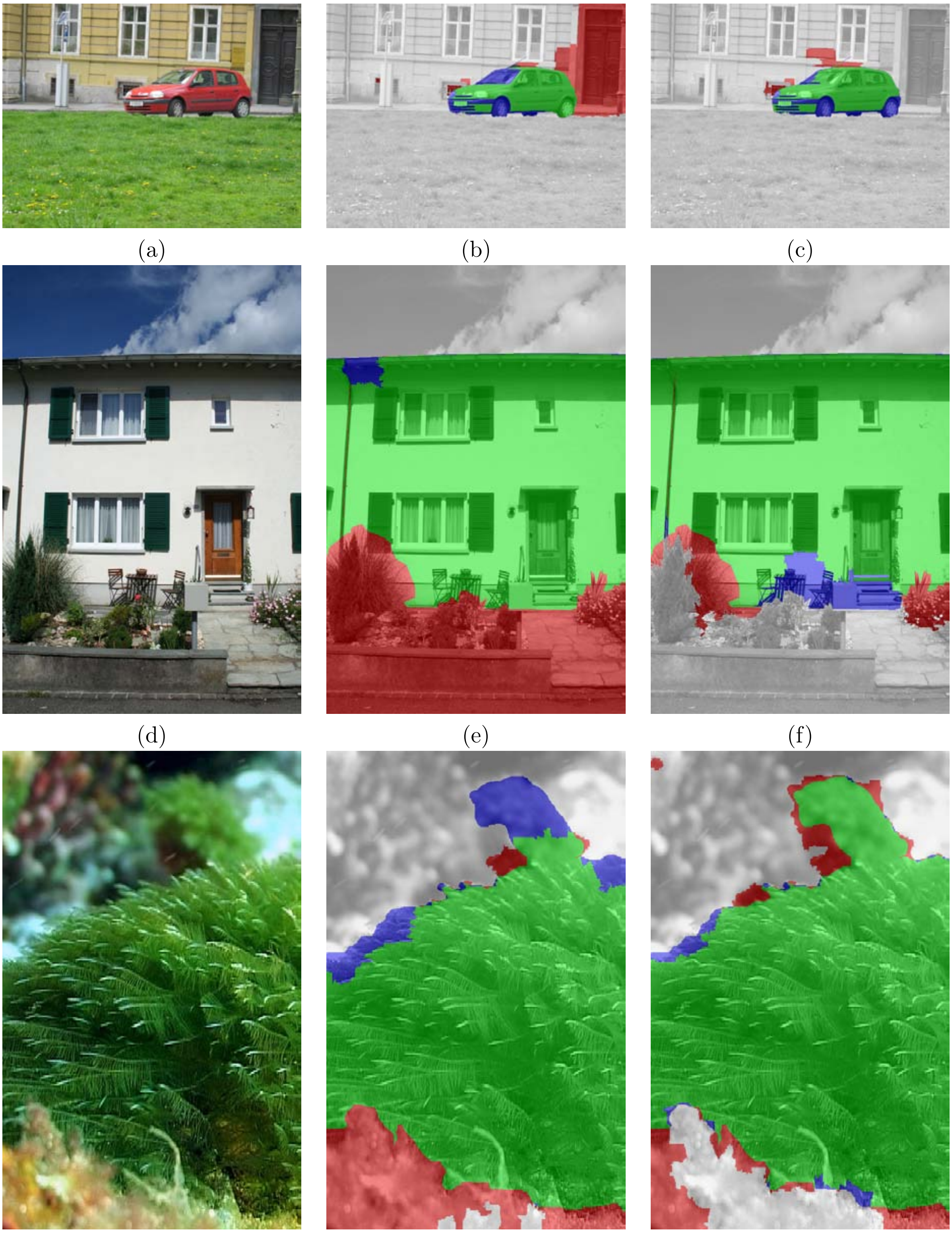


Figure 4.13.: Results of Graz-02 cars [Opelt et al., 2006], eTRIMS [Korč & Förstner, 2009] building facades, and algae with standard CRF (b, e, h) and with ISC-CRF (c, f, i).

$\mathbf{Q}_{ll}$, more precisely, corresponding variances σ^2. Its inverse $\mathbf{P} = \mathbf{Q}_{ll}^{-1}$ acts as weight matrix in the objective function (Eq. 3.15), putting high weights on accurate observations and lower weights on less accurate ones. A well justified choice of σ is important to gain reliably adjusted height values. Standard deviations σ of observations are summarized in table 4.7. It should be noted that in case of approximate standard deviation values (e.g., σ_{Alt}), conservative values are taken in order to avoid overconfident estimation.

σ_{Opt}	σ_{Alt}	$\sigma_{db,proj}$	$\sigma_{db,slant}$	σ_{lay}	σ_{InSAR}	σ_{ρ}	σ_{θ}
0.37	0.5	0.74	0.39	0.78	0.5	0.01	0.02

Table 4.7.: Standard deviations σ of observations: in the orthophoto σ_{Opt}, sensor altitude σ_{Alt}, double-bounce lines projected to orthophoto geometry $\sigma_{db,proj}$, double-bounce lines in slant range geometry $\sigma_{db,slant}$, layover edge in single SAR magnitude image σ_{lay}, InSAR height σ_{InSAR}, sun elevation angle σ_{ρ}, and incidence angle of the SAR sensor σ_{θ}; all standard deviations in unit meter except σ_{ρ} and σ_{θ} (in degree).

The optical image used for experiments is an orthophoto, it has been projected from camera space to ground using a digital terrain model. σ_{Opt} of observations in the orthophoto depends on how accurately object positions (e.g., shadow, building) are mapped. It is described by the image resolution, defined as the shortest distance between two objects still allowing to distinguish them. Point spread functions (PSF) can be used to derive an approximate resolution value if knowing the quantization interval (i.e., ground sampling distance (GSD), size of a pixel on ground). A PSF describes how a single point in object space is mapped in image space. Assuming a single object spreading across multiple pixels, such oversampling can be used to determine a resolution of an image. Point-like objects with high contrast to background are rare in the orthophoto, shadow edges are used instead. Gradient profiles of the optical intensity channel at multiple shadow edges, uniformly distributed over the entire image, are recorded. A Gaussian function is fitted to each profile and its standard deviation in unit pixel is used as a scaling factor of the GSD. Averaging over all edge measurements across the image results in a factor of 1.19 pixels. Multiplying with the GSD gives a resolution of $\sigma_{Opt} = 1.19 \times 0.31\ m = 0.3689\ m \approx 0.37\ m$. All profile measurements and processing is done with the software EDGE [Jacobsen, 2009]. Standard deviation σ_{Opt} is assigned to all observations conducted in the orthophoto (e.g., distances between roof edges $\mathbf{e}$ and nadir $\boldsymbol{\nu}$).

Sensor altitudes are needed in equations 3.8 & 3.10. The corresponding standard deviation σ_{Alt} mainly depends on GPS altitude accuracy of the airplane. In order to take into account all potential inaccuracies, a rather conservative value of $\sigma_{Alt} = 0.5\ m$ is chosen.

Height measurement 3.8, being a function of the InSAR double-bounce line that is projected to orthophoto geometry, calls for $\sigma_{db,proj}$ in addition to σ_{Opt} and σ_{Alt}. Standard deviation $\sigma_{db,proj}$ represents the positioning accuracy of the projected double-bounce line. Instead of a theoretical derivation, $\sigma_{db,proj}$ is estimated empirically. Lines at buildings where the corresponding edge $\mathbf{w}$ (cf. Fig. 3.9) is well visible in the orthophoto, too, are investigated. In an ideal case, double-bounce line and the edge where building walls meet ground should exactly match. In order to determine deviations, distances orthogonally to double-bounce lines are measured to corresponding edges $\mathbf{w}$. Double-bounce line positions vary between minus and plus two pixels, $\sigma_{db,proj}$ is set to $\sigma_{db,proj} = 2 \times \sigma_{Opt} = 0.74\ m$.

Gable roof heights are computed in slant range geometry. Double-bounce line and edges of the bright single-bounce line are observed. It is shown in [Wegner et al., 2010] that double-bounce lines can be automatically segmented with an accuracy of about one pixel in slant range. This value is chosen as standard deviation for, both, double-bounce lines and edges of single-bounce lines in slant range $\sigma_{db,slant} = 0.39\ m$, which corresponds to the resolution in range direction.

Standard deviation σ_{lay} is needed for the near range end of layover in a SAR magnitude image in ground range geometry, as used for height computation in equation 3.10. An investigation of a straight building roof edge, which should also map exactly straight in the SAR magnitude image, shows deviations of plus an minus two pixels leading. This frayed signature of a roof edge in layover can be seen in figure 2.7(a), for example. Thus, two times the SAR range resolution is considered as standard deviation $\sigma_{lay} = 0.78\ m$.

The accuracy of maximum robust InSAR heights within layover ramps of flat roof buildings is assessed by direct comparison to LiDAR ground truth. Heights vary about ± 0.5 meters around reference heights, a standard deviation $\sigma_{InSAR} = 0.5\ m$ is set. It corresponds to Aes-1 InSAR height accuracy values reported by Schwäbisch & Moreira [1999].

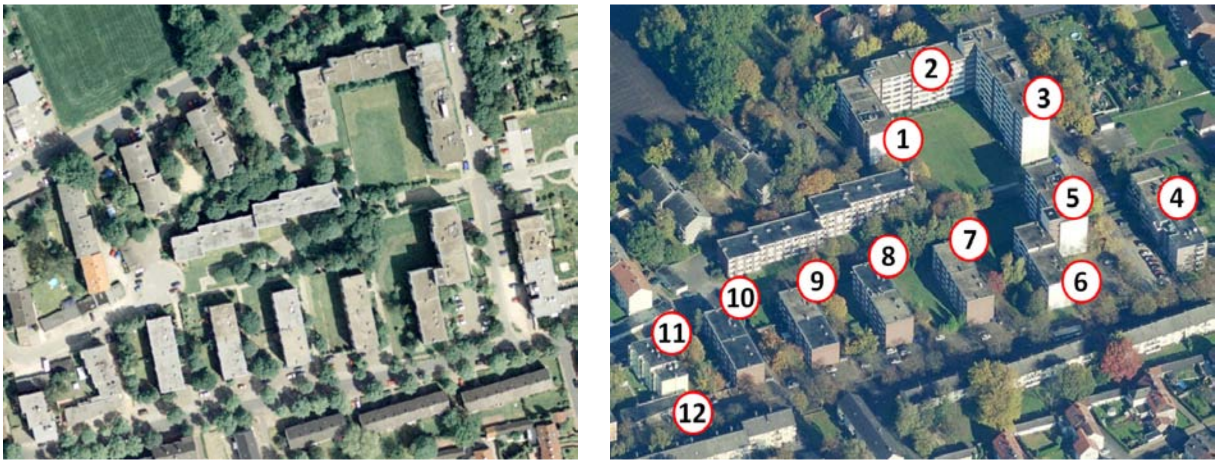
Incidence angle observations of sun and viewing angle observations of the SAR sensor are made, too. Sun incidence angle ρ is computed with software NREL SOLPOS with a standard deviation $\sigma_\rho = 0.01°$ [Rymes, 2000]. SAR sensor viewing angle θ at the location of a particular gable roof building in the image is assumed to have a standard deviation of $\sigma_\theta = 0.02°$.

4.3.2. Comparison of different height measures

Results of building height measurements proposed in section 3.2.1 are presented here. Outcomes of single measurements defined in equations 3.7 to 3.12 are shown. Adjusted results combining single measurements per building using *least squares adjustment with functionally dependent parameters* will be provided in the following section. In this section, focus is, first, on flat roof buildings and, second, on gable roof buildings.

All three-dimensional visualizations compare measured heights to ground truth acquired with airborne laserscanning (LiDAR). Buildings are embedded into a digital terrain model[3]. Brownish roof tops indicate the deviation with respect to the LiDAR reference height. Plus and minus symbols denote whether a measured building height is too low (minus) or too high (plus). A white building colour indicates that the corresponding height measurement could not be performed due to missing observations in data. It should be noted that building footprints have been extracted manually from

[3] The DTM originates from the same LiDAR data as building height references. Terrain heights in figures are slightly exaggerated for visualization purposes.



(a) (b)

Figure 4.14.: Flat roof buildings (a) in the orthophoto used for measurements and (b) oblique view from Microsoft Bing Maps^{TM} (© 2011 Microsoft Corporation, © 2010 Blom).

the orthophoto. Visualization with footprints automatically detected with CRF classification would hamper interpretability due to irregular footprint shapes. A refinement of building detection results, either through an improved classification or via post-processing, would allow using automatically detected footprints. It is left for future work.

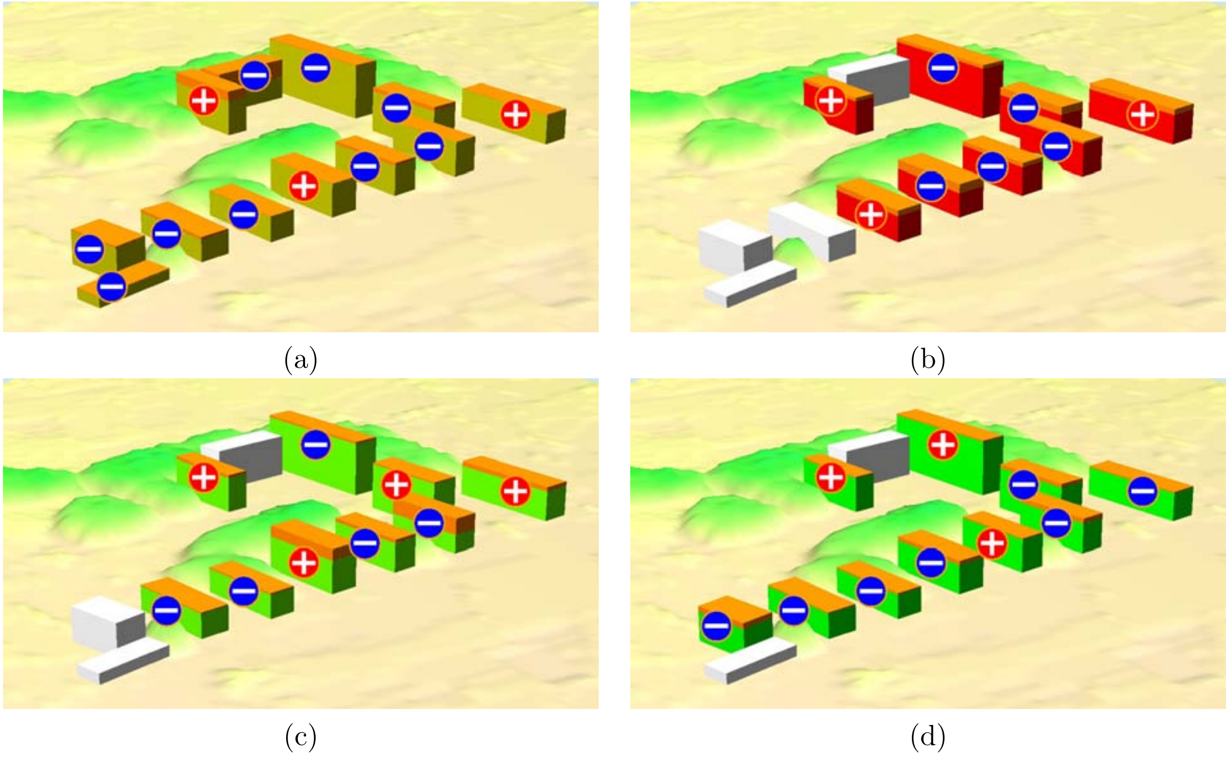
All cut-outs of a SAR magnitude image (entire image given in Fig. 4.2) of the InSAR image pair are shown in ground range geometry in order to ease interpretability and comparisons with corresponding orthophoto cut-outs. The latter are part of the original orthophoto provided in figure 4.14. Indices of buildings are overlaid to oblique optical images taken as screenshots from Microsoft Bing Maps^{TM}. Their perspective was chosen to correspond as much as possible to the three-dimensional visualizations in order to facilitate understanding.

Flat roof buildings

All height measurements of flat roof buildings are based on observations in the orthophoto, a SAR magnitude image, or InSAR heights. Several primitives are segmented manually in order to prepare for observations: Roof edges $\mathbf{e}$, nadir point $\boldsymbol{\nu}$, shadow edge $\mathbf{s}$, and the location $\mathbf{w}$ where building walls meet ground in the orthophoto[4].

SAR double-bounce lines $\mathbf{db}$ are automatically segmented and projected to orthophoto geometry. Robust maximum InSAR heights h_l in layover ramps are extracted automatically in slant range geometry, too, within parallelograms as explained in section 3.2.1. The widths of parallelograms is set as a function of the maximum unambiguous building height in InSAR data of approximately 30 meters. In the Dorsten scene no buildings higher than 30 meters occur. In case of a lower

[4] Automatic segmentation of all primitives is possible, too, but calls for adapted processing decreasing artefacts as much as possible. It is left for future work. The goal here is to assess the best achievable building height accuracy and thus manual measurements are well justified.



(a)

(b)

(c)

(d)

Figure 4.15.: Results of flat roof building height measurements determined via (a) *sun shadow* (h_s), (b) *perspective distortion* in the optical image (h_{pd}), (c) *double-bounce line* and roof edge overlap (h_{db}), (d) *SAR layover* (h_l).

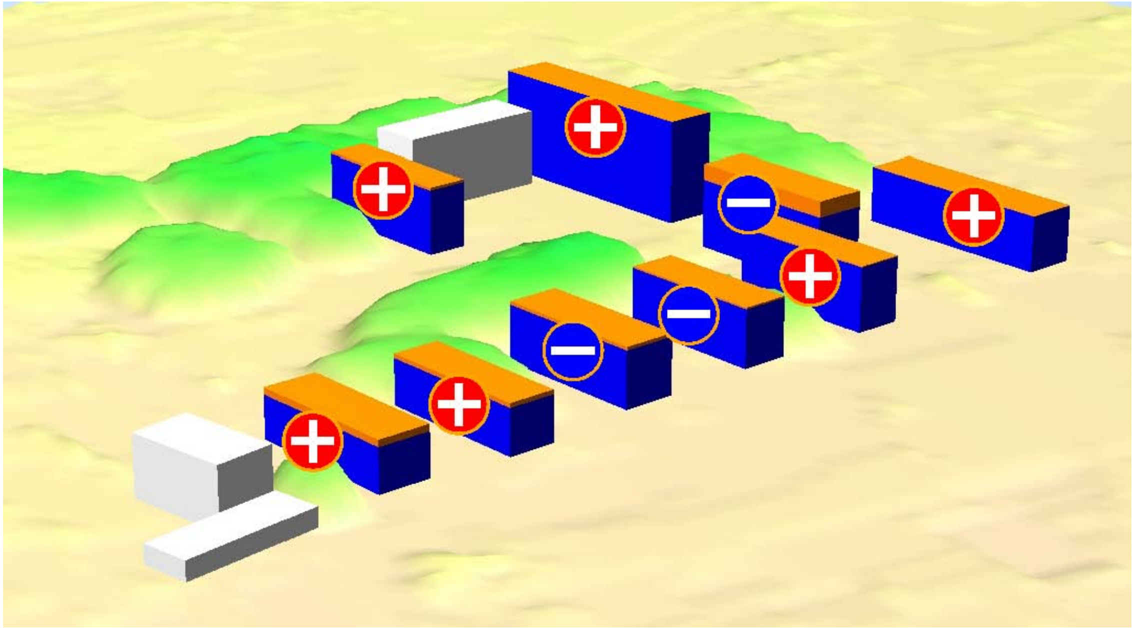


Figure 4.16.: Results of flat roof building height measurements determined via robust maximum InSAR height in layover phase ramp (h_{InSAR}).

maximum phase to height ambiguity or higher buildings phase unwrapping has to be conducted before computing h_l. It is unnecessary here.

Observations are conducted automatically between manually or automatically segmented primitives, which are all lines (except the nadir point $\boldsymbol{\nu}$ in the orthophoto). Multiple measurements along line primitives are done, paying attention to uncorrelated values. Two observations at a line primitive have to be separated by a minimum distance corresponding to the resolution of the data (cf. Tab. 4.7).

An oblique optical image (screenshot from Microsoft Bing Maps$^{\mathrm{TM}}$) in figure 4.14(b) gives an impression of the three-dimensional extent of flat roof buildings used for testing. Indices are assigned to each building being part of tests concerning the proposed height measures and adjustment. The corresponding part in the orthophoto is shown in figure 4.14(a), magnitude SAR image and InSAR heights can be viewed in figure 3.10. All single height measurements per flat roof building are given in table B.1 in Annex B, a diagram providing an overview is given in figure 4.21.

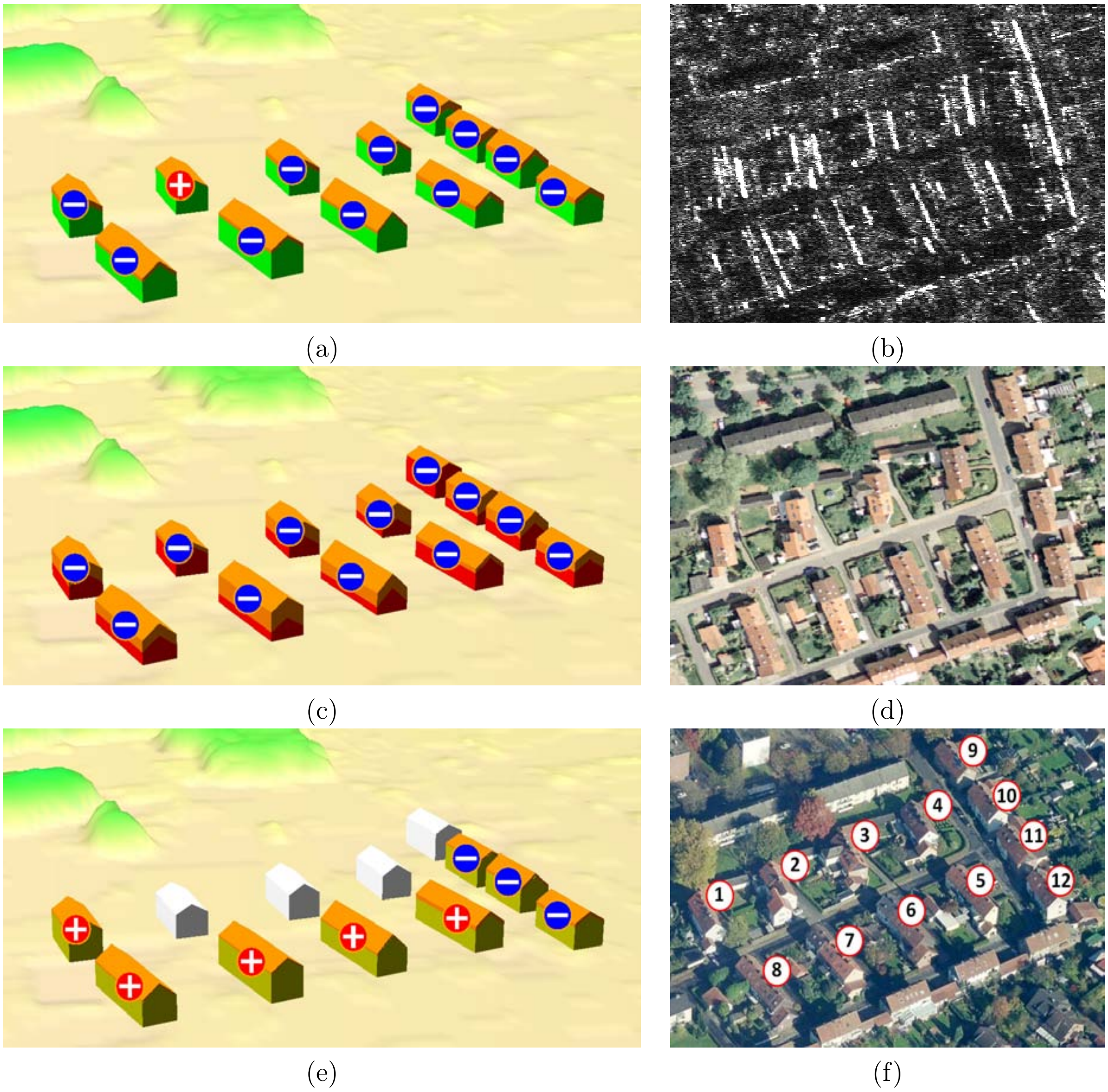
Figure 4.15 illustrates three-dimensionally single measured heights without an explicit height value as input to the equation. All measures are purely based on inherent height information caused by characteristic effects. In figure 4.16 robust maximum InSAR heights h_{InSAR} measured in layover phase ramps are visualized.

Gable roof buildings

Multiple gable roof buildings are investigated in terms of height measurements. Often, not all proposed height measurements can be conducted due to missing observations in data. All single height values per gable roof building are given in table B.2 in Annex B.

No robust maximum InSAR heights in layover phase ramps (h_{InSAR}) are measured for gable roof buildings. Similar to the intensity values, all phase contributions of the tilted roof plane, from eave up to ridge, are collected in the near range bright line. The phase value of the dominant scatterer on the roof plane, anywhere between eave and ridge, is recorded. Layover between double-bounce and single-bounce line originates from the building facade, the highest point in this layover phase ramp does not correspond to the roof ridge (i.e., the building height), but to the eave. Moreover, gable roof buildings are usually smaller than flat roof buildings and thus signal from elevated objects in front of them like trees interferes. As a consquence, h_{InSAR} is not a good measure for gable roof building heights, it is not conducted. Height measurements of gable roof buildings are based on observations in the orthophoto, one SAR magnitude image, and double-bounce lines. As in case of flat roof buildings, some primitives are segmented manually: Building width c, roof edges $\mathbf{e}$, nadir point $\boldsymbol{\nu}$, shadow edge $\mathbf{s}$, the location $\mathbf{w}$ where building walls meet ground in the orthophoto, near and far range edges of bright SAR lines caused by single-bounce reflection at the tilted roof plane.

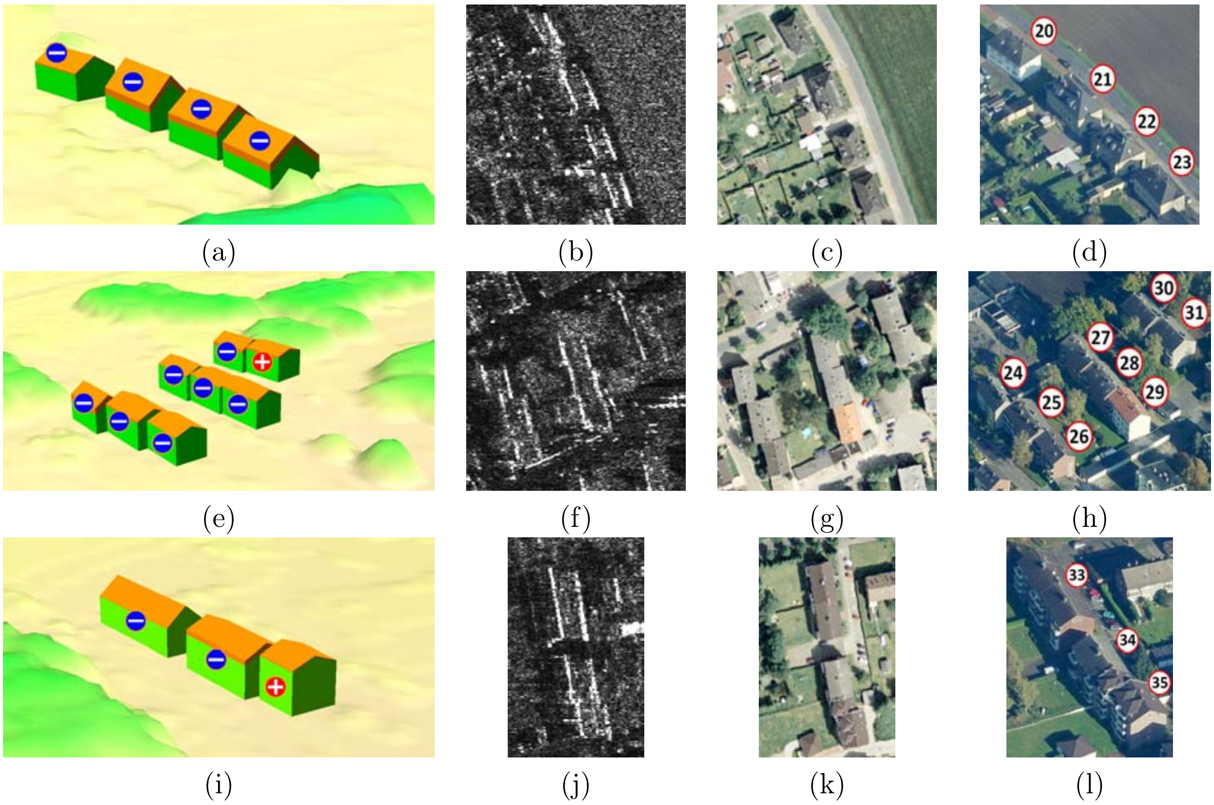
SAR double-bounce lines $\mathbf{db}$ are automatically segmented and projected to orthophoto geometry. Observations (e.g., distance measurements) are conducted automatically between manually or



(a) (b) (c) (d) (e) (f)

Figure 4.17.: Height measurements for gable roof buildings 1 to 12: (a) via *parallel bright lines* in SAR data (h_r), Eq. 3.11 & 3.12), (b) corresponding cut-outs of one SAR magnitude image of the InSAR image pair (ground range geometry, range direction right to left), (c) optical *perspective distortion* heights (h_{pd}), Eq. 3.8 with $\mathbf{w}$ instead of $\mathbf{db}$), (d) corresponding cut-out of the orthophoto, (e) heights via *sun shadow* of roof ridge (h_s), Eq. 3.7), (f) oblique view from Microsoft Bing Maps$^{\text{TM}}$ (© 2011 Microsoft Corporation, © 2010 Blom).

automatically segmented line primitives. Like in case of flat roof buildings, attention is payed to achieve uncorrelated values. Distances between single observations either correspond to orthophoto or InSAR resolution, 0.37 meters or 0.385 meters, respectively. The mean of each observation is introduced into the height measurement equation. Single building height measurements of gable roof buildings one to twelve are shown in figure 4.19. Corresponding cut-outs of SAR magnitude image (Fig. 4.19)(b), orthophoto (Fig. 4.19(d)), and oblique optical image from Microsoft Bing

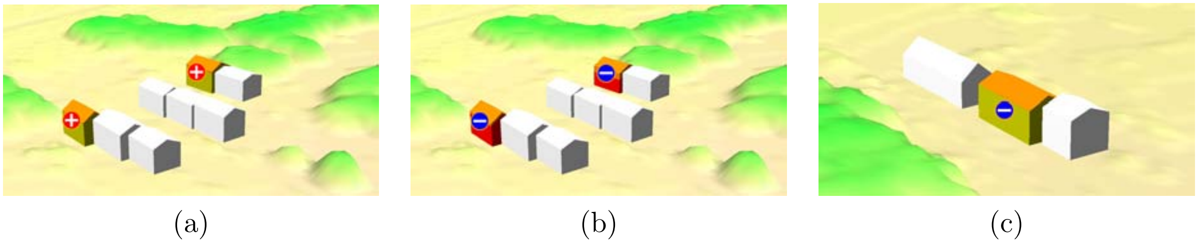


(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 4.18.: Results of building height measurements for gable roof buildings via *parallel bright lines* in SAR data (h_r, Eq. 3.11 & 3.12), corresponding cut-outs of one SAR magnitude image of the InSAR image pair (ground range geometry, range direction right to left), of the orthophoto, and oblique view from Microsoft Bing Maps$^{\mathrm{TM}}$ (© 2011 Microsoft Corporation, © 2010 Blom): (a)-(d) buildings 20 to 23, (e)-(h) buildings 24 to 31, (i)-(l) buildings 33 to 35.



(a) (b) (c)

Figure 4.19.: Results of building height measurements for gable roof buildings via *sun shadow* of roof ridge (h_s, Eq. 3.7) and *optical distortion* (h_{pd}): (a) *sun shadow* heights 24 to 32, (b) *optical distortion* heights 24 to 32, (c) *sun shadow* heights 33 to 35; no heights can be determined at buildings depicted in white; no heights can be measured at buildings 20 to 23 (cf. Fig. 4.18(a)-(c)).

Maps$^{\mathrm{TM}}$ (Fig. 4.19(f)) are provided, too. Three different height measurements could be conducted for most buildings: h_r via parallel bright lines in SAR data, h_{pd} via optical perspective distortion heights at northern short building sides (cf. Fig. 4.19(d)), and h_s via sun shadow of the roof ridge. At buildings 20 to 23 no heights based on the extent of sun shadow can be determined because, dealing with hip roofs, the ridge is invisible in the shadow (cf. Fig. 4.18(c)). Moreover, optical distortion cannot be used for height measurements neither due to buildings being positioned very close to nadir. Only very few heights depending on sun shadow (Eq. 3.7) and optical distortion (Eq. 3.8 with w instead of db) can be obtained in building groups 24 to 35. Most buildings in figure 4.19 appear white, no additional heights besides the ones depending on parallel bright SAR lines (cf. Fig. 4.18) can be determined. In total, merely two gabel roof buildings facilitate computation of three different heights (cf. Fig. 4.18(d) & Fig. 4.19(a,b)), one building with two heights is present (cf. Fig. 4.18(g) & Fig. 4.19(c)).

4.3.3. Adjusted building heights

Least squares adjustment with functionally dependent parameters as introduced in section 3.2.2 is used to estimate a single height per building. Observations are weighted according to values for σ provided in section 4.3.1. An inital height h_0 is needed for adjustment, here the most accurate height measurement is taken as first approximation h_0. All single height values per flat roof building are given in table B.1 and those of gable roof buildings in B.2 in Annex B.

Two different adjustments are processed for each flat roof building, the first ($h_{b,noI}$) combining all heights except InSAR heights h_{InSAR} and the second (h_b) including h_{InSAR}, too. The first adjustment $h_{b,noI}$ signifies how accurately building height estimation can be conducted relying merely on inherent data effects without any explicit height like h_{InSAR}. Figure 4.20(a) illustrates $h_{b,noI}$ three-dimensionally. Results of h_b combining all measured heights of a building are shown in figure 4.20(b). All measured flat roof building heights can be compared in the diagram in figure 4.21. Posterior standard deviations $\hat{\sigma}_b$ (cf. 3.19) are calculated for every adjusted building height h_b. They are a precision measure, indicating the interior accuracy of height adjustment. Differences of adjusted heights h_b to LiDAR refence heights $\Delta_{b,L}$ represent the absolute accuracy attained via least squares adjustment. A comparison of $\hat{\sigma}_b$ and $\Delta_{b,L}$ is given in diagram 4.25.

Adjusted gable roof heights are shown in diagrams 4.22(b) & 4.24. They are visualized three-dimensionally in figures 4.22(a) & 4.23. Corresponding posterior standard deviations $\hat{\sigma}_b$ and absolute differences to LiDAR refence heights $\Delta_{b,L}$ are given in diagram 4.26.

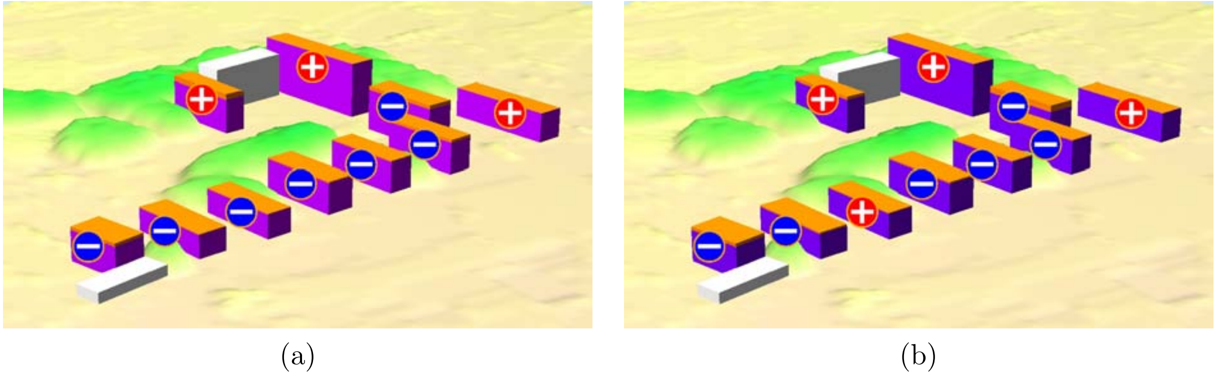


Figure 4.20.: Results of flat roof building heights after *least squares adjustment with functionally dependent parameters* combining all possible height measures at each building: (a) without InSAR heights ($h_{b,noI}$), (b) with InSAR heights (h_b); no adjustment is conducted at buildings two and twelve (coloured white) due to only a single height measurement.

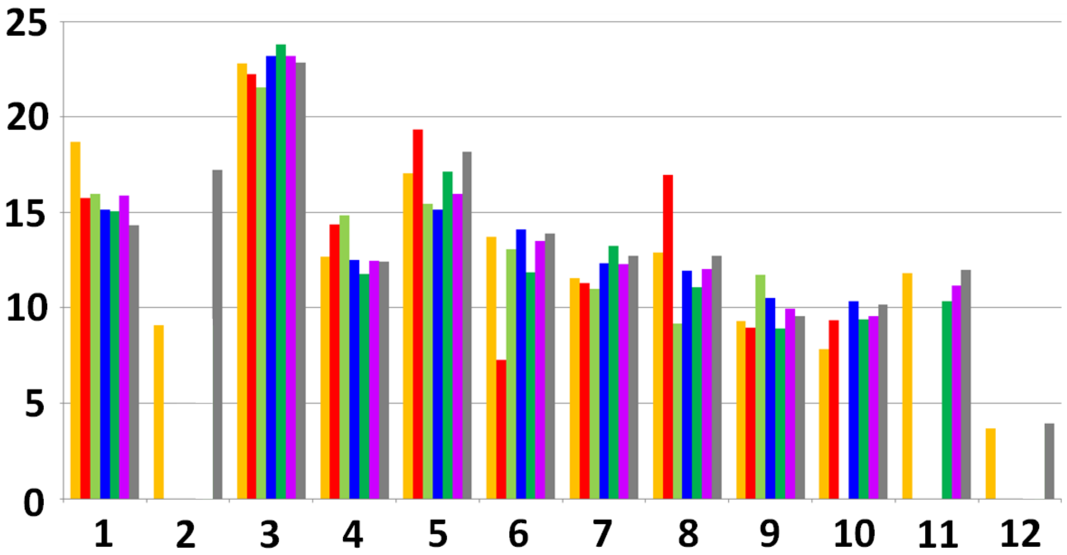
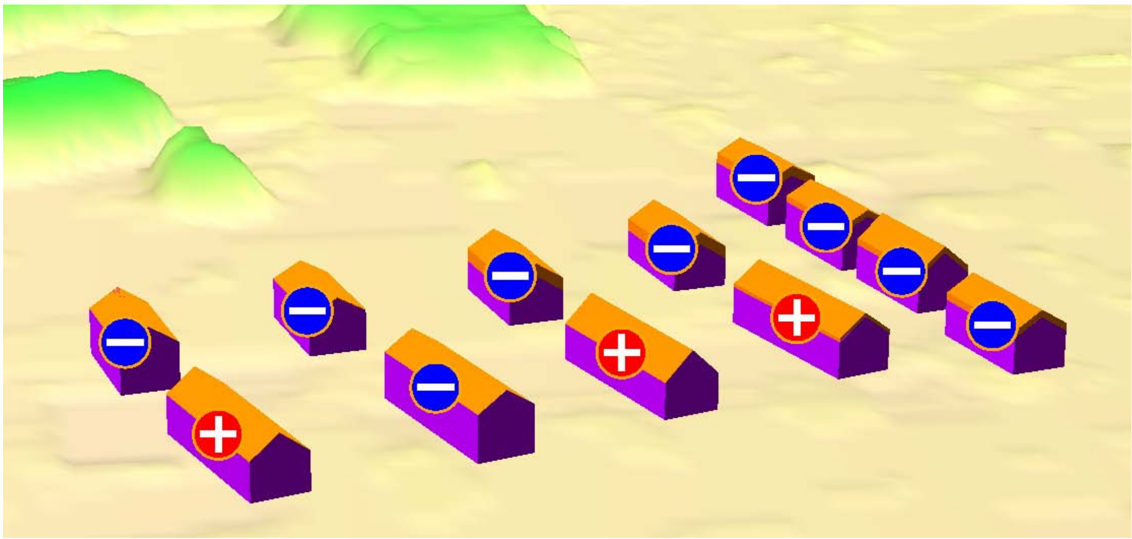
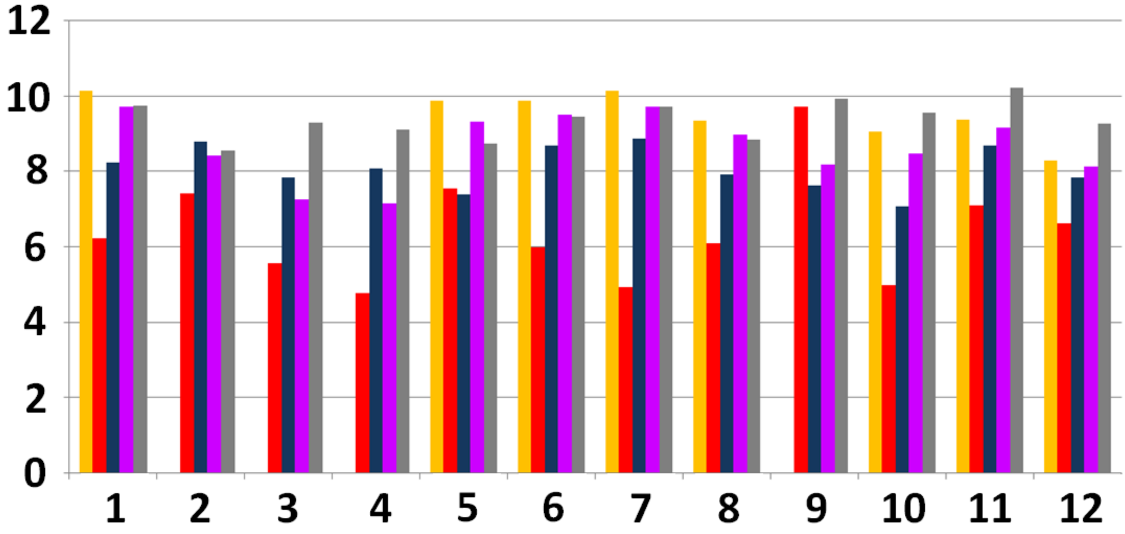


Figure 4.21.: Flat roof building heights 1 to 12 measured via h_s *sun shadow* (yellow), h_{pd} optical *perspective distortion* (red), h_{db} overlap of roof edge and *double-bounce line* (light green), h_{InSAR} robust maximum *InSAR heights* in layover ramp (blue), h_l *layover* in SAR magnitude image (dark green), and h_b all possibilities combined with *least squares adjustment* (purple) compared to the *LIDAR reference heights* (grey); abscissa: building indices as in figure 4.14(b), ordinate: height above ground in unit meter; note: absent heights indicate that no heights could be measured due to missing observations.



(a)



(b)

Figure 4.22.: Adjusted gable roof heights of buildings 1 to 12: (a) three-dimensional plot; (b) height diagram: h_s via *sun shadow* (yellow), h_{pd} optical *perspective distortion* (red), h_r *parallel SAR lines* (dark blue), h_b all possibilities combined with *least squares adjustment* (purple) compared to h_L *LIDAR reference heights* (grey); abscissa: building indices as in figure 4.19(f), ordinate: height above ground in unit meter; note: absent heights indicate that no heights could be measured due to missing observations.

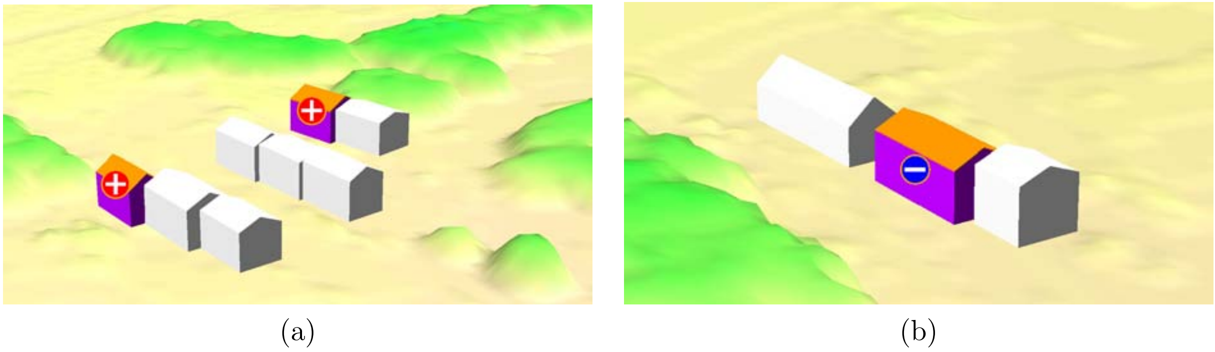


Figure 4.23.: Results of *adjusted gable roof heights*: (a) adjusted heights h_b of buildings 24 to 32 and (b) of 33 to 35; note: only observations of buildings 24, 30, and 34 allowed for more than one measurement and an adjustment (cf. Fig. 4.19).

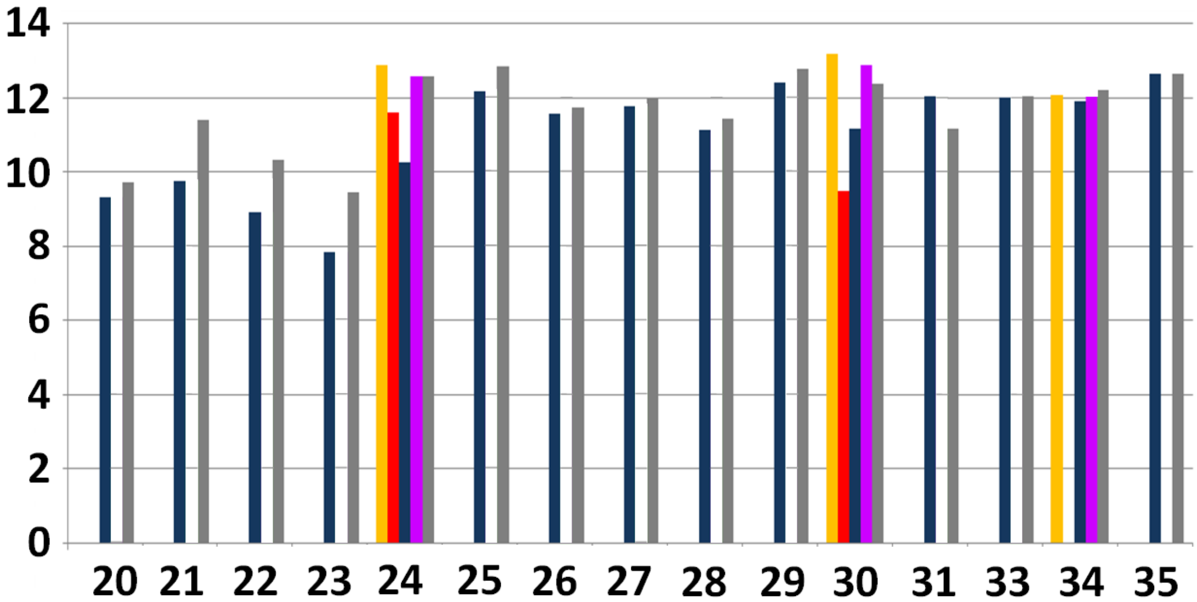


Figure 4.24.: *Adjusted gable roof heights* of buildings 20 to 31 and 33 to 35: h_s via *sun shadow* (yellow), h_{pd} optical *perspective distortion* (red), h_r *parallel SAR lines* (dark blue), h_b all possibilities combined with *least squares adjustment* (purple) compared to h_L *LIDAR reference heights* (grey); abscissa: building indices as in figures 4.18(f,h,l), ordinate: height above ground in unit meter; note: absent heights indicate that no heights could be measured due to missing observations.

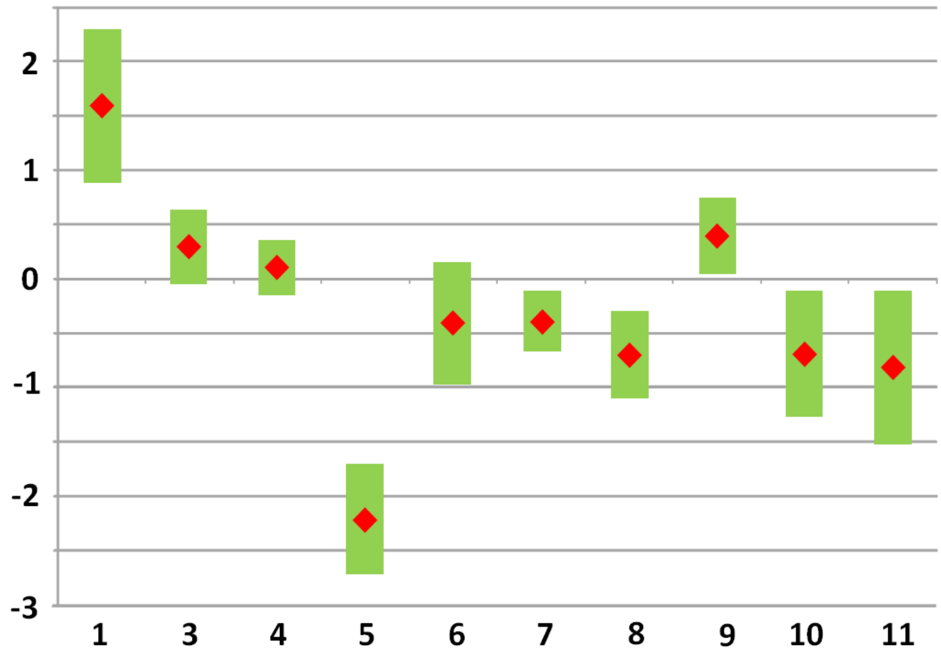


Figure 4.25.: Flat roof building *height deviations*: *Posterior standard deviations* $\hat{\sigma}_b$ (green) and *absolute differences* $\Delta_{b,L}$ to LiDAR reference (red); abscissa: only buildings where an adjustment could be conducted (more than one measured height) are shown, ordinate: deviation around the reference height in unit meter, zero corresponds to the LiDAR reference height.

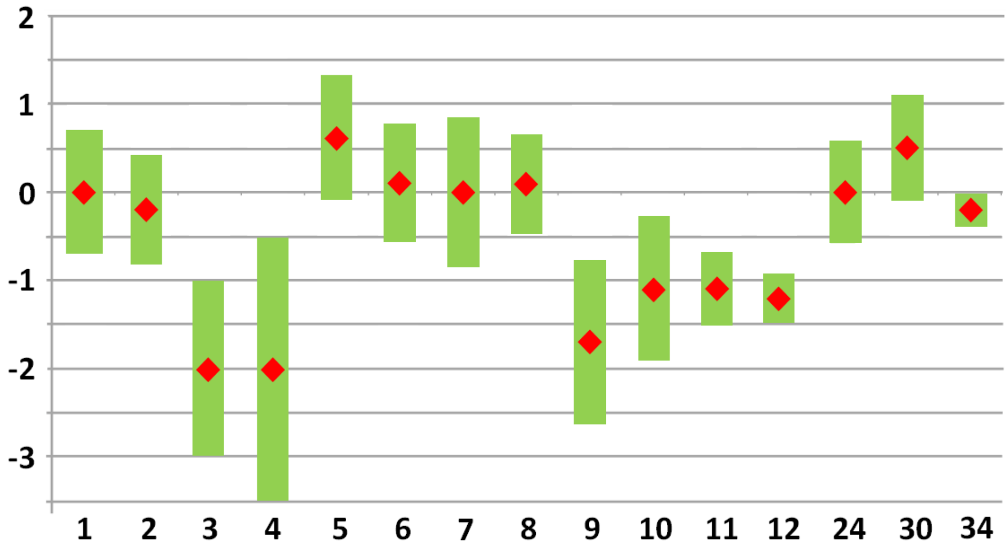


Figure 4.26.: Gable roof building *height deviations*: *Posterior standard deviations* $\hat{\sigma}_b$ (green) and *absolute differences* $\Delta_{b,L}$ to LiDAR reference (red); abscissa: only buildings where an adjustment could be conducted (more than one measured height) are shown, ordinate: deviation around the reference height in unit meter, zero corresponds to the LiDAR reference height.

5. Discussion

In this chapter experimental results of the previous chapter are discussed. Novel methods introduced in Chapter 3 are critically evaluated based on those tests, particular characteristics are explained, conclusions are drawn, and potential improvements suggested. The first section discusses building detection results (5.1) using the CRF and its different context formulations. In the second section (5.2), height estimation results are interpreted.

5.1. Assessment of CRF building detection results

This section discusses building detection results achieved with different CRF formulations. First, the general method is compared to Maximum Likelihood (ML) and Markov Random Field (MRF). Then, results of region-based CRF, gradient discontinuity constraint, impact of SAR double-bounce lines, and ISC-CRF are explained.

CRF results versus ML and MRF

Maximum Likelihood classification achieves a much lower TPR, but a significantly reduced FPR, too, compared to the standard CRF, both, on patch-graphs (Tab. 4.1, Fig. 4.5(a)-(c)) and on region-graphs (Tab. 4.2, Fig. 4.5(d)-(f)). It correctly detects 70.7% of all building patches, only misclassifying 16.1% of non-building patches. Compared to ML, the CRF yields a much higher detection rate (TPR 77.6 vs. 70.7), but a higher FPR (23.4 vs.16.1), too. A visual comparison of ML and CRF outcomes on patches (cf. Fig. 4.5(a)-(c) & Fig. 4.7(a)-(c)) and regions (Fig. 4.5(d)-(f) & Fig. 4.7(d)-(f)) reveals different object detection characteristics. ML detects most buildings, but sometimes only partially, whereas the CRF detects almost the entire building. If comparing ML and CRF results on object level, the CRF detects more patches (or regions) per building. This is an advantage of the CRF and important for further processing on object level. False positives of ML occur at places where no building at all is located (e.g., street in upper half of Fig. 4.5(a)). The majority of CRF false positives occurs between closely located buildings, a drawback of the CRF. This result shows that the CRF tends to over-smooth although observations are included in the prior.

One possible explanation is need for much more training data in case of the CRF. For ML only very few parameters, variances and covariances of the multi-variate Gaussians, have to be trained. The

CRF requires automatic tuning of more than 100 parameters after quadratic expansion of feature vectors with the same amount of data. Significantly more annotated training data is thus needed compared to ML.

The MRF shows the highest TPR (88.4% & 86.0%) with the drawback of a significantly increased FPR (31.4% & 26.8%). This high FPR compared to such of ML already hints at the principle drawback of MRFs if dealing with building detection in urban areas: It works well as long as buildings are sparsely distributed, but fails at dense building groups. MRF classification leads to strong over-smoothing effects between buildings and to false positives far away from any building (Fig. 4.5(g)-(i)). TPR (88.4 vs. 77.6) and FPR (31.4 vs. 23.4) increase significantly compared to the CRF on patches, the same applies both methods on regions. It is due to the simple Ising model of the MRF prior. Unlike CRFs, MRFs do not consider observations $\mathbf{x}$ in the prior (cf. Eq. 2.12 & Eq. 2.15). Only labels are compared regardless of edge features. This strong smoothing of the MRF leads to almost all parts of a building being detected as such. But gaps between neighbouring buildings are misclassified. The MRF is not suitable for the detection of single buildings because it detects entire settlement areas (which could be an application).

In conclusion, the MRF is not suitable at all, but good performance of ML highlights some room for improvement considering the CRF. Over-smoothing is the CRF's main drawback calling for more training data and more sophisticated contextual modelling.

Performance evaluation of region-based graphs

An irregular graph based on image regions improves building detection results compared to the regular patch graph, but not as much as one could expect. Although regions of an over-segmentation well preserve object boundaries, the CRF does not succeed in discriminating both classes in feature space as anticipated. The most likely reason is that the feature distributions do not sufficiently well characterize and distinguish building from non-building nodes. More sophisticated and distinctive features have to be computed and introduced to the CRF classification framework.

The major advantage of the region-based CRF compared to the patch-based CRF is a significantly reduced computation time. A graph of regions enables a more than ten times faster computation due to less nodes and edges (cf. Tab. 4.3).

A multi-scale segmentation only marginally improves results compared to a single segmentation at the highest scale. Most probably this is due to rather simple multi-scale integration via the feature vectors. Features of coarser scales are simply concatenated with those of higher scales. However, this way of a multi-scale incorporation does not allow for coarse regions to significantly extent across buildings because this would lead to mixtures of feature distributions of buildings and background. Segmentation parameters as given in section 4.2.2 generate the coarsest segmentation with regions still preserving building boundaries (cf. Fig. 4.6). An advantage of concatenating features of different scales is that a more complex three-dimensional graph struture is avoided. The size of the graph stays unchanged, computation time is only marginally increased due to more features and thus more weights to be trained. Multi-scale adaption without changing the graph leads to a smoothing effect inside objects as long as regions of the coarsest scale do not significantly violate object boundaries. It leads to a slight improvement of the TPR (80.7% vs. 79.5%) compared to the single-scale segmentation due to less missed building superstructures, for example, chimneys in roof areas.

A graph based on image regions should be chosen if the segmentation algorithm well respects boundaries of the objects of interest. It is much faster in terms of computation time compared to the standard patch graph, more expressive context formulation and learning is possible. If segmentation does not appropriately respect object shapes, a patch graph with small patches should be preferred.

Gradient-based discontinuity constraint

The discontinuity constraint as function of the mean gradient between two neighbouring regions has only a small impact on building detection. Both, TPR (78.4% vs. 79.5%) and FPR (20.8% vs. 22.0%) are slightly decreased compared to a standard CRF. A reason is that gradients do not only occur between building regions and their direct neighbours, but at other elevated objects, too. Trees lead to strong gradients, too, for example.

This constraint could work better if the testing scene would contain only large flat roof buildings in an industrial zone without other elevated objects. One possibility to improve classification for the Dorsten scene would be to introduce a multi-class CRF. All object categories being separated from others by high gradients could then be captured within single classes, building and elevated vegetation, for example.

Another reason possibly hampering performance is the fact that gradients between elevated objects and sun shadow only occur at one side of the object (depending on the sun azimuth). If introducing the gradient discontinuity constraint as a function of direction, too, and considering all elevated objects as separate classes, improvements may be achieved.

Significance of SAR double-bounce line

Comparing building detection results with (Fig. 4.7(d)-(f)) and without double-bounce line features (Fig. 4.9), no significant difference can be recognized. TPR (79.1 % versus 79.5 %) and FPR (21.9 % versus 22.0 %) are on the same level. Features derived from SAR lines have a negliable impact on classification results, which is somehow surprising reconsidering figure 4.4, for example.

In order to explain potential reasons, it should be noted that double-bounce lines have a particular character as building hint. In case lines are present, it is very likely that pixels behind them (in range direction) belong to buildings. However, absent lines do not indicate that a building is unlikely because double-bounce lines do not occur at all buildings due to obstructions, for example. As a consequence, the classifier does not learn double-bounce line features to be distinctive for building detection; they consequently receive very low weights and have almost no impact on the shape of the decision surface in feature space.

A second reason is the rather simple linear model used as discriminative function to model the decision surface between both classes, building and background, in feature space (Eq. 2.17 & 2.18). Although a quadratic expansion of feature space takes place leading to a quadratic decision surface in original feature space, this seems to be insufficient. Two major ways of resolving this task exist.

First, the building class can explicitly be subdivided in two classes, one of buildings with double-bounce lines and one without. We would than have to learn and infer three distinct classes in total with the CRF classifier: two building classes and one background class. From a feature space perspective, this would mean introducing an additional decision surface to distinctively separate buildings and background.

A second possibility is to keep classification binary with one decision surface in feature space, but constructing a decision surface of higher order than quadratic. It can be achieved in two different ways: First, the linear model of equations 2.17 & 2.18 could directly be replaced by a function of higher order, a quadratic or kubic polynomial, for example; second, the linear model is kept, but the kernel mapping function, currently quadratic, is replaced by one of higher order. As with Support Vector Machines, various kernel mapping functions could be evaluated to introduce a higher dimensional feature space, keeping the original linear model function in the enlarged feature space, but turning it into a higher order one in original feature space. All possibilities will be investigated in future work[1].

Although the double-bounce line does not make a significant contribution to building detection, due to reasons aforementioned, it is essential for height estimation as will be discussed in section 5.2.

ISC-CRF applied to simulated data

Applied to the simulated scene (Fig. 4.12)(a) the ISC-CRF significantly reduces the FPR from 6.8% to only 0.8%. As edge feature vector the standard CRF considers the absolute difference of adjacent node feature vectors in order to support or suppress smoothing. Since grey buildings and grey streets are located very closely in colour feature space, the standard CRF cannot well distinguish those two object categories, neither based on node features nor on edge features. It leads to some street regions being miscassified as building (Fig. 4.12)(b). The ISC-CRF learns the arrangement of subcategories "street" and "grey building" (besides all other subcategories) implicitly and is thus

[1] Nonetheless, it should be noted that, both, introducing an additional class or a more sophisticated discriminative decision surface, lead to more parameters to be trained. Additional parameters need more observations calling for an enlarged training database, which results in higher computation costs (in terms of time and memory). Concerning a direct introduction of a polynomial discriminative function, training would become more challenging. A more complex log-likelihood objective function would call for adapted approaches during optimization, first and second derivatives have to be redefined.

able to discriminate the two. Such being the case, streets are correctly classified as background, although original colour features are not distinctive (Fig. 4.12)(c). This result shows that scenes with context of low complexity can benefit from implicit scene context.

Results of varying numbers of cluster centers are presented in figure 4.11(a). The ISC-CRF FPR varies about 1 % (from 0.8 % to 1.8 %) and no significant trend is observable. Changing the number of k-means cluster centers has a very small impact on classification performance, but of course on computation time. A rather small number of cluster centers is beneficial. Segmentation scale is adapted to each scene separately (and context radii are a function of the mean region size) because it depends on the scales of context and objects. This makes the ISC-CRF highly flexible and easy to adapt to new scenes.

Results of noise experiments based on the simulated scene are depicted in figure 4.11(b). The FPR of the ISC-CRF stays below that of the standard CRF at all noise levels. Furthermore, the ISC-CRF is slightly more robust to noise because its FPR starts increasing later (approx. 90 % vs. approx. 80 %).

Experiments with simulated data show that the general concept of implicit scene context helps discriminating object classes if original features are not distinctive enough. It is robust to noise, even more robust than the standard CRF, and changing the currently manually adjusted number of cluster centers has only a small impact on results.

Application of ISC-CRF to object detection in real data

Building detection combining optical and double-bounce line features is neither significantly improved nor deteriorated applying the ISC-CRF (cf. Tab. 4.5, Fig. 4.12 & 4.7(d)-(f)). Compared to a standard CRF, the FPR is slightly reduced from 22.0% to 21.2%, but the TPR decreases (78.1% vs. 79.5%), too.

Possible reasons explaining this outcome are, first, the highly complex context of the urban scene and, second, need for more training data. Spatial arrangements of subcategories show a significant variation, which could not sufficiently well learned by the ISC-CRF. Cluster patterns of buildings and surrounding subcategories of the background class are very diverse. This high diversity of leads to no significant pattern being learned. A first attempt to improve results could be to use much more training data.

Application of the ISC-CRF to other data besides remote sensing imagery underlines transferability of the method. All three scenes (cars, facades, and algae) have context of medium complexity. In all three cases the ISC-CRF decreases the FPR significantly in comparison to the standard CRF (cf. Tab. 4.6).

The highest decrease of the FPR (7.3 % vs. 22.1%) is achieved with building facades (Fig. 4.13(d)-(f)). Only using the standard CRF, colour and gradient features do not sufficiently well discriminate a building facade from foreground (Fig. 4.13(e)). However, incorporating implicit scene context (based on the same colour and gradient features), the CRF can well distinguish building facade from vegetation and doorway in the foreground (Fig. 4.13(f)). In contrast to the simulated urban scene of Fig. 4.12(a), for example, the facade data shows a different object and context structure. The facade appears only once and covers a very large area of the test image. It is not entirely surrounded by context, but only above and below. A similar distribution of object and contextual subcategories occurs in alga images. Nonetheless, algae are smaller in comparison to image size and their boundaries are frayed. Again, implicit scene context decreases the FPR (23.7 % vs. 37.0 %) while increasing the TPR (84.5 % vs. 75.7 %) (cf. Fig. 4.13(g) & (i)). Cars have a much smaller size relative to the entire image and they are completely surrounded by context. Nonetheless, subcategories in this context show a certain ordering (although not as distinct as in case of building facades): cars usually appear on roads, often buildings are in the background, but vegetation may appear all around the car. A lower FPR is achieved if incorporating implicit scene context in the CRF (4.3 % vs. 16.4 %), while the TPR is slightly increased (88.1 % vs. 86.6 %). This can also be seen comparing the improved results in figure 4.13(c) to those generated with the standard CRF (Fig. 4.13(b)). Moreover, computation time using the implicit scene context potential does only marginally increase by several seconds per image.

In conclusion, implicit scene context significantly improves object detection if applied to scenes with context of medium and low complexity (Fig. 4.12 & 4.13). Remote sensing data proves to be the most challenging classification task because context has the highest degree of complexity. Building detection is not significantly improved with the ISC-CRF, the implicit scene context concept has to be formulated in a more sophisticated way (ideas in Chapter 6).

5.2. Evaluation of estimated building heights

First, results of single height measurements of flat and gable roof buildings are discussed, second, adjusted heights are evaluated.

Single height measurements

The *sun shadow* generally delivers accurate results h_s as long as the assumption of locally flat terrain is not violated. It cannot be used for height measurements if neighbouring objects like trees hamper recognizability of shadow edges. In case sun elevation angles of close to ninety degrees occur, shadow cannot be recognized at all and no heights may be determined.

Perspective distortion, caused by the central perspective of the camera, is a good measure h_{pd} if certain conditions are met. Buildings should be located far away from nadir and high buildings can better be measured than smaller ones. Buildings with flat roofs are preferred because observations can be made along the entire roof edge. Heights of gable roof buildings can only be calculated at the two endpoints of the roof ridge. However, all heights h_{pd} of gable roof buildings one to twelve are measured systematically too low, which can be seen in figure 4.19(c). Reconsidering the corresponding cut-out of the orthophoto in figure 4.19(d), only northern short building sides can be used to measure because the edge where building wall meets ground has to be recognized. The northern roof ridge endpoint is not located in the same plane as the building wall, but slightly juts out orthogonally by approximately half a meter. Due to very short distances between endpoint of roof ridge and building wall meeting ground in the orthophoto, this effect leads to a systematic underestimation of building heights.

Overlap of roof edge and *double-bounce line* delivers satisfying results h_{db} (cf. Fig. 4.15(c)) if some conditions are met. Basically, the same restrictions as previously discussed for perspective distortion apply. In addition, the positioning accuracy of the double-bounce line projected to orthophoto geometry is crucial. Obstructions of building facades can lead to absent or only partially present lines. Moreover, erroneous InSAR height values lead to lines being displaced in orthophoto geometry. This projection error propagates to height determination, which happens at flat roof building eight (Fig. 4.15(c)), for example.

The *layover* length of a flat roof building signature in a SAR magnitude image enables reliable height measurements h_l (Fig. 4.15(d)) in case the near range edge can be recognized. A stronger layover effect, evoked by smaller SAR sensor viewing angle θ and at higher buildings, is benefitial, but limitations apply in urban areas due to interfering signal from adjacent elevated objects.

Robust maximum *InSAR heights* h_{InSAR} deliver satisfying results if all assumptions made in section 3.2.1 are validated. Widths of parallelograms are subject to the constraint that all buildings in the scene are smaller than the maximum unambiguous height. Phase unwrapping prior to calculating h_{InSAR} becomes necessary (and adapation of the parallelogram width) if higher buildings occur or the InSAR baseline is chosen smaller.

Determining heights h_r of gable roof buildings via the *distance between double-bounce and single-bounce lines* in SAR magnitude images [Thiele et al., 2007b, 2010a] works sufficiently well under the condition that both lines can be recognized (Fig. 4.19(a) & 4.18(a,e,i)). Width b of the single-bounce line is usually very narrow thus complicating observations.

In general, all single height measurement possibilities underly constraints. Not all measurements can be conducted at each building. Sometimes, only a single height can be determined (e.g., gable roof buildings 27 to 29). In case multiple measurements are possible, least squares adjustment with functionally dependent parameters can combine and weight all single height measures.

Adjusted heights

Including InSAR heights h_{InSAR} into adjustment slightly improves results, but adjustment without them leads to good results, too (cf. Fig. 4.20(a) & (b)). This comparison shows that based merely on inherent height information in optical and SAR data, building heights can be estimated with a reasonable accuracy.

Comparing posterior standard deviations $\hat{\sigma}_b$ and absolute deviations of estimated heights from the LiDAR reference $\Delta_{b,L}$ (cf. Tab. B.1 & B.2) shown in figures 4.25 & 4.26, many differences are outside the interval of plus or minus a single standard deviation. It indicates that some of the posterior standard deviations, which represent height precision, tend to deliver too confident values. One interpretation would be the existence of systematic errors. Similar height measurements thus lead to an expectation value with an offset from the true height. However, offsets of flat roof buildings in the same image do neither have equal absolute values nor do they occur in the same direction (i.e., no equal algebraic signs). This is different with gable roof buildings, most heights are estimated too low, which is most probably due to parallel SAR line heights h_r and sun shadow heights h_s (cf. Fig. 4.19(a,c)). Shadow heights are too small because distances between ridge and shadow are measured too short. The reason is that the roof ridges jut out.

Another explanation would be too small a priori standard deviations σ for some of the observations. Their inverse values are contained as weights in matrix $\mathbf{P}$, which is a factor of equation 3.19 that computes the a posteriori standard deviation. In general, the very small size of the stochastic sample (i.e., number of different height measurements per building) with a maximum of five values for some flat roof buildings is a problem. Future work will comprise an investigation of this issue.

Absolute differences $\Delta_{b,L}$ of measured heights to reference heights represent the accuracy. Heights of eight out of ten flat roof buildings have been estimated with an accuracy better than one meter (Fig. 4.26). The worst accuracy is estimated for building five, with a height -2.2 meters below the reference. In case of gable roof buildings, the maximum height offset is smaller (-2.0 meters below the reference at buildings three and four), but more heights are estimated with an accuracy worse than one meter. Furthermore, all of them are systematically too low (Fig. 4.25). One explanation would be that heights h_r obtained via parallel bright lines and particularly shadow heights h_s are all underestimated (cf. Fig. 4.19(a,c)) as outlined previously.

Sun shadow height h_s contributes most to adjusted height h_b because corresponding standard deviations of observations are lower than those of other height measurements. It thus gains a high influence on the final adjusted building height and leads to inaccuracies if measured wrong as already mentioned. Thus, the a priori standard deviation of observations included in shadow height measurements have to be carefully revised in future work if dealing with gable roof buildings. All other height measurement possibilities are approximately on the same accuracy level if considering optimum conditions (e.g., high building far away from nadir in case of h_{pd}).

In general, least squares adjustment with functionally dependent parameters proves to be a valuable tool.

6. Conclusions and future work

This chapter concludes this thesis. The most important methodological insights are highlighted and ideas for future work are presented. It will first be dealt with contextual object detection before turning to building height estimation.

Contextual object detection

The first objective as stated in section 1.1 was to *develop an innovative solution for the detection of buildings in urban areas merging information derived from one high-resolution SAR acquisition and one optical image.* It was achieved by the introduction of a *contextual probabilistic approach learning its parameters* within a Conditional Random Field framework. Besides building detection, the CRF proved to be *generally applicable to object detection in any scene* and to outperform MRFs. Training of parameters obviates the need for software changes in terms of parameter tuning and re-definition of rules. Moreover, probabilistic learning unveils high-dimensional patterns in feature space that could most probably not be detected manually by human experts.

Replacing the regular graph of image patches with an irregular graph of image regions significantly accelerates computation. It better preserves object boundaries and captures the topology of a scene enabling distinct contextual learning. The concept of implicit scene context enhances classification of data with context of low and medium complexity. Learning contextual links between subcategories of semantically annotated object classes is able to uncover underlying patterns in an unsupervised manner. Integration of a multi-scale segmentation into the CRF via concatenation of feature vectors improves classification avoiding much increase of computation time. The original graph structure stays unchanged, standard learning and inference techniques can be applied.

Although CRFs provide good results, the principal drawback of over-smoothing could not completely be resolved concerning building detection. Edge features, as currently generated, do not sufficiently suppress smoothing via the prior. The gradient discontinuity constraint in its present form did not change this. Improvements could be achieved by introducing, first, more discriminative edge features, and, second, by re-designing the prior energy term. Furthermore, much larger training datasets would facilitate distinct learning of patterns significantly. In the following, some ideas are presented that may inspire future research efforts. A multi-scale segmentation integrated explicitly into the graph resulting in a three-dimensional structure, where messages are passed between regions of neighbouring scales, too (e.g., [Kohli et al., 2008, 2009]), would establish new ways of sophisticated contextual learning. Object feature distributions and contextual links could be captured separately at different scene scales, too. For example, four scales could be distinguished if dealing with building detection: building parts (roof planes, chimneys), single buildings (shadow, front yards), building blocks (characteristic pattern of streets, trees, building rows), and settlement (inner city cores versus suburban areas). In addition, completely representing a scene topology in multiple scales with a graph would enable inter-scale contextual learning. Region-ancestry concepts as suggested by Lim et al. [2009] could be included and re-formulated in a CRF.

The ISC-CRF does not learn highly complex urban context of object class subcategories appropriately. One idea is to consider the shapes of regions for context histogram ranges. Instead of simply drawing circular ranges around the region centroid, one could enlarge the original region, keeping its shape, by certain ranges. Elongated street regions, for example, sticking out of the first circular range and being counted twice (again in the second range), would be extended by the same distance in any direction thus avoiding double counting. Circular ranges reach out further into the image perpendicularly to an elongated region, with respect to its boundaries, than lengthwise. Introduction of shape would avoid this bias and give equal importance to any direction.

Another idea would be to turn implicit scene context into a partially explicit scene context descriptor. A multi-class CRF could be designed, training data still being only partially labeled, but with more than two different classes. In remote sensing data, for example, one could semantically annotate buildings, streets, grassland, and high vegetation. All remaining classes would be contained in a background class. Context histograms could then learn characteristic patterns of explicitly labeled classes as well as of unsupervised clustered subcategories of the background class. Furthermore, hidden subcategories could be formulated probabilistically as latent variables within a Hidden Conditional Random Field as suggested by Quattoni et al. [2007].

SAR double-bounce lines did not significantly improve classification, which is, first and foremost, due to the simple linear discriminant function of the CRF. More sophisticated functions should be introduced, either directly into the CRF or via feature space mapping with kernel functions of higher order than quadratic. It will lead to much more accurate decision surfaces, capable of better adaption to training data, thus improving results in general and reducing the over-smoothing effect.

In general, the CRF prior has not been used to explicitly learn contextual relations of object categories, yet. It basically has stayed a smoothing term, where smoothing degree is tuned. Furthermore, only local to regional context has been learned, yet, although the CRF allows for global context learning. One idea would be to use large cartographic databases, for example Open Street Map or ATKIS©, to train global contextual relations between urban objects like roads, buildings, and vegetated areas. Learning this global context would be rather fast because cartographic data already exists in vector format. We could exploit very large cartographic databases in a relatively short time. Instead of only determining one-by-one relations of the node of interest to a neighboring node we could think of detecting particular context constellations. The basic idea is that certain groups of objects are hints for nearby buildings, for example. However, as soon as we move away from pair-wise relations towards comparisons of more than two nodes, we have to adapt training and inference [Kohli et al., 2009]. Another promising possibility for learning context in cartographic databases is *Graph Matching*. It has lately been deployed to handwriting recognition and to object recognition in imagery. Applying Graph Matching to global context training in order to support building detection seems to have a great potential although combinatorical issues will arise. Instead of relying solely on Graph Matching, it could be integrated via an additional potential into a CRF. The association potential of the CRF framework would then learn local object features, the interaction potential regional context, and global patterns in cartographic data could be learned via Graph Matching. These directions of thought will be focussed on in future research.

Height estimation

Reconsidering the second objective, *accuracy assessment of building height estimation based on a single SAR acquisition and one optical image*, *least squares adjustment with functionally dependent parameters* led to buildings heights of meter accuracy. For the first time, multiple height measurement possibilities of such fused data are combined in a *sound stochastic framework* and and *jointly adjusted*. Posterior standard deviations act as *measure of precision* facilitating to *judge reliability of height estimates*. Achieved height accuracies can be viewed as the best possible with the data configuration at hand.

Concepts for building height estimation could potentially be used as prior knowledge to facilitate phase unwrapping in urban areas. For example, one could think of introducing double-bounce lines in front of buildings as discontinuity constraint during phase unwrapping, which assumes smooth surfaces without any height jumps in state-of-the-art algorithms. Furthermore, an optical image could be added to InSAR data to estimate initial rough building heights with the concepts outlined in section 3.2. Those initial height guesses could potentially serve as prior knowledge if the height accuracy is better than the 2π phase unwrapping disambiguity.

Conclusion

Both objectives of this thesis have been met successfully. Probabilistic contextual object detection, learning its parameters, is a highly useful tool for a wide range of applications. Conditional Random Fields in particular provide great flexibility for contextual classification with a single comprehensive probabilistic framework. Nonetheless, more in-depth research has to deal with this topic. Instead of suppressing a smoothing term, learning of explicit contextual relations between object categories is needed. The simple linear discriminant function is to be replaced by a higher order model, either through direct formulation in the CRF energy term, via feature space mapping, or both. Over-smoothing would potentially be avoided and SAR double-bounce lines could be recognized as essential building hints by the classifier. With respect to object detection in remote sensing data, big semantically annotated training datasets should be established to unleash the full power of learning techniques. They could serve as benchmark to ease comparisons between different methods, too.

Bibliography

Ball GH, Hall DJ (1967) A clustering technique for summarizing multivariate data. *Systems Research and Behavioral Science*, 12 (2): 153–155.

Bamler R, Hartl P (1998) Synthetic aperture radar interferometry. *Inverse Problems*, 14 (4): 1–54.

Bartholomew-Biggs M (2008) *Nonlinear Optimization with Engineering Applications.* Berlin: Springer.

Belongie S, Malik J, Puzicha J (2002) Shape Matching and Object Recognition Using Shape Contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24 (24): 509–522.

Benediktsson JA, Swan PH, Ersoy OK (1990) Neural Network Approaches Versus Statistical Methods in Classification of Multisource Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*, 28 (4): 540–552.

Bennett AJ, Blacknell D (2003) The Extraction of Building Dimensions from High Resolution SAR Imagery. In: *IEEE International Radar Conference*: 182–187.

Benz UC, Hofmann P, Willhauck G, Lingenfelder I, Heynen M (2004) Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. *ISPRS Journal of Photogrammetry and Remote Sensing*, 58 (3-4): 239–258.

Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, B: 192–236.

Biederman I, Mezzanotte RJ, Rabinowitz JC (1982) Scene Perception: Detecting and Judging Objects Undergoing Relational Violations. *Cognitive Psychology*, 14: 143–177.

Bishop CM (2006) *Pattern Recognition and Machine Learning.* Berlin: Springer.

Bolter R (2003) *Buildings from SAR: Detection and Reconstruction of Buildings from Multiple View High Resolution Interferometric SAR Data.* PhD thesis, Institut für Maschinelles Lernen und Darstellen, Technische Universität Graz.

Bolter R, Leberl F (2000) Detection and Reconstruction of Human Scale Features from High Resolution Interferometric SAR Data. In: *IEEE International Conference on Pattern Recognition*: 291–294.

Boykov Y, Veksler O, Zabih R (2001) Fast Approximate Energy Minimization via Graph Cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23 (11): 1222–1239.

Briem GJ, Benediktsson JA, Sveinsson JR (2002) Multiple Classifiers Applied to Multisource Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*, 40 (10): 2291–2299.

Brunner D, Bruzzone L, Ferro A, Lemoine G (2009) Analysis of the Reliability of the Double Bounce Scattering Mechanism for Detecting Buildings in VHR SAR Images. In: *IEEE Radar Conference*

Brunner D, Lemoine G, Bruzzone L (2008) Building Height Retrieval From Airborne VHR SAR Imagery Based On An Iterative Simulation And Matching Procedure. In: Michel U, Civco DL, Ehlers M, Kaufmann HJ (eds) *SPIE: Remote Sensing for Environmental Monitoring, GIS Applications, and Geology VIII.* SPIE, 7110: 291–296.

Brunner D, Lemoine G, Bruzzone L, Greidanus H (2010) Building Height Retrieval From VHR SAR Imagery Based on an Iterative Simulation and Matching Technique. *IEEE Transactions on Geoscience and Remote Sensing*, 48 (3): 1487–1504.

Campbell JB (2002) *Introduction to remote sensing.* The Guilford Press, 3rd edition.

Carbonetto P, de Freitas N, Barnard K (2004) A Statistical Model for General Contextual Object Recognition. In: Pajdla T, Matas J (eds) *ECCV 2004, LNCS 3021* (pp. 350–362). Berlin: Springer.

Cellier F, Colin E (2006) Building height estimation using fine analysis of altimetric mixtures in layover areas on polarimetric interferometric X-band SAR images. In: *IEEE International Geoscience and Remote Sensing Symposium*: 3987–3990.

Cellier F, Oriot H, Nicolas J (2006) Hypothesis Management for building reconstruction from high resolution InSAR imagery. In: *IEEE International Geoscience and Remote Sensing Symposium*: 3639–3642.

Clifford P (1990) Markov random fields in statistics. In: Grimmett G, Welsh D (eds) *Disorder in Physical Systems: A Volume in Honour of John M. Hammersley* (pp. 19–32). Oxford University Press.

Comaniciu D, Meer P (1999) Mean shift analysis and applications. In: *IEEE International Conference on Computer Vision*

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

Cour T, Benezit F, Shi J (2005) Spectral Segmentation with Multiscale Graph Decomposition. In: *IEEE International Conference on Computer Vision and Pattern Recognition*

Crowther PS, Cox RJ (2005) A Method for Optimal Division of Data Sets for Use in Neural Networks. In: et al. RK (ed) *Lecture Notes in Computer Science* (pp. 1–7). Berlin: Springer.

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

Dare P, Dowman I (2000) A new approach to automatic feature based registration of SAR and SPOT images. In: *International Archives of Photogrammetry and Remote Sensing*. ISPRS Symposium Amsterdam 2000, 33.

Denis L, Tupin F, Darbon J, Sigelle M (2009) Joint Regularization of Phase and Amplitude of InSAR Data: Application to 3-D Reconstruction. *IEEE Transactions on Geoscience and Remote Sensing*, 47 (11): 3774–3785.

Dong Y, Forster B, Ticehurst C (1997) Radar backscatter analysis for urban environments. *International Journal of Remote Sensing*, 18 (6): 1351–1364.

Ehlers M, Tomowski D (2008) On Segment Based Image Fusion. In: Blaschke T, Lang S, Hay GJ (eds) *Object-Based Image Analysis: Spatial Concepts for Knowledge-Driven Remote Sensing Applications, Lecture Notes in Geoinformation and Cartography* (pp. 735–754). Berlin: Springer.

Ferretti A, Prati C, Rocca F (2000) Nonlinear Subsidence Rate Estimation Permanent Scatterers in Differential SAR Interferometry. *IEEE Transactions on Geoscience and Remote Sensing*, 38 (5): 2202–2212.

Franceschetti G, Iodice A, Riccio D (2002) A Canonical Problem in Electromagnetic Backscattering From Buildings. *IEEE Transactions on Geoscience and Remote Sensing*, 40 (8): 1787–1801.

Franceschetti G, Iodice A, Riccio D, Ruello G (2003) SAR Raw Signal Simulation for Urban Structures. *IEEE Transactions on Geoscience and Remote Sensing*, 41 (9): 1986–1995.

Frey BJ, MacKay DJ (1998) A Revolution: Belief Trees: Belief Propagation in Graphs With Cycles. In: Jordan MI, Kearns MJ, Solla SA (eds) *Advances in Neural Information Processing Systems.* Cambridge, MA: MIT Press.

Fulkerson B, Vedaldi A, Soatto S (2009) Class Segmentation and Object Localization with Superpixel Neighborhoods. In: *IEEE International Conference on Computer Vision*

Galleguillos C, Belongie S (2010) Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114: 712–722.

Galleguillos C, McFee B, Belongie S, Lanckriet G (2010) Multi-Class Object Localization by Combining Local Contextual Interactions. In: *IEEE Conference on Computer Vision and Pattern Recognition*

Gamba P, Houshmand B (2000) Hyperspectral and IFSAR Data for 3D Urban Characterization. In: *IEEE International Geoscience and Remote Sensing Symposium.* 6: 2611–2613.

Gould S, Fulton R, Koller D (2009) Decomposing a Scene into Geometric and Semantically Consistent Regions. In: *IEEE International Conference on Computer Vision*

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

Guiasu S, Shenitzer A (1985) The Principle of Maximum Entropy. *The Mathematical Intelligencer*, 7 (1): 42–48.

Guida R, Iodice A, Riccio D (2010) Height Retrieval of Isolated Buildings From Single High-Resolution SAR Images. *IEEE Transactions on Geoscience and Remote Sensing*, 48 (7): 2967–2979.

Guida R, Iodice A, Riccio D, Stilla U (2008) Model-Based Interpretation of High-Resolution SAR Images of Buildings. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1 (2): 107–119.

Hanssen RF (2001) *Radar Interferometry: Data Interpretation and Error Analysis.* In: Freek van der Meer: *Remote Sensing and Digital Image Processing.* Kluwer Academic Publishers.

He W, Jäger M, Reigber A, Hellwich O (2008) Building Extraction from Polarimetric SAR Data using Mean Shift and Conditional Random Fields. In: *European Conference on Synthetic Aperture Radar.* 3: 439–443.

He X, Zemel RS, Carreira-Perpiñán MA (2004) Multiscale Conditional Random Fields for Image Labeling. In: *IEEE Conference on Computer Vision and Pattern Recognition*

Hégarat-Mascle SL, Bloch I, Vidal-Madjar D (1997) Application of Dempster-Shafer Evidence Theory to Unsupervised Classification in Multisource Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 35 (4): 1018–1031.

Heitz G, Koller D (2008) Learning Spatial Context: Using Stuff to Find Things. In: Forsyth D, Torr P, Zisserman A (eds) *ECCV 2008, LNCS 5302* (pp. 30–43). Berlin: Springer.

Hendrix EM, G.-Toth B (2010) *Introduction to Nonlinear and Global Optimization*. Berlin: Springer.

Hepner GF, Houshmand B, Kulikov I, Bryant N (1998) Investigation of the Integration of AVIRIS and IFSAR for Urban Analysis. *Photogrammetric Engineering and Remote Sensing*, 64 (8): 813–820.

Hill RD, Moate CP, Blacknell D (2006) Urban Scene Analysis from SAR Image Sequences. In: *European Conference on Synthetic Aperture Radar*

Hoberg T, Rottensteiner F (2010) Classification of settlement areas in remote sensing imagery using Condtional Random Fields. In: *IntArchPhRS*. 38 (7A).

Hoberg T, Rottensteiner F, Heipke C (2010) Classification of Multitemporal Remote Sensing Data using Condtional Random Fields. In: *6th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS 2010), Istanbul, Turkey*

Hoiem D, Efros AA, Hebert M (2008) Putting Objects in Perspective. *International Journal of Computer Vision*, 80: 3–15.

Hong TD, Schowengerdt RA (2005) A Robust Technique for Precise Registration of Radar and Optical Satellite Images. *Photogrammetric Engineering and Remote Sensing*, 71 (5): 585–593.

Inglada J, Giros A (2004) On the Possibility of Automatic Multisensor Image Registration. *IEEE Transactions on Geoscience and Remote Sensing*, 42 (10): 2104–2120.

Jacobsen K (2009) EDGE. software supplement of the Bundle Block Adjustment Leibniz University Hannover (BLUH) system developed at the Institute of Photogrammetry and GeoInformation (IPI), Germany.

Jahangir M, Blacknell D, Moate CP, Hill RD (2007) Extracting information from shadows in SAR imagery. In: *IEEE International Conference on Machine Vision*: 107–112.

Klonowski J, Koch KR (1997) Two Level Image Interpretation Based on Markov Random Fields. In: Förstner W, Plümer L (eds) *Semantic Modeling for the Acquisition of Topografic Information from Images and Maps* (pp. 37–55). Basel: Birkhäuser.

Kluckner S, Bischof H (2010) Image-based building classification and 3D modeling with super-pixels. In: *IntArchPhRS*. 38: 233–238.

Koch H, Pakzad K, Tönjes R (1997) Knowledge Based Interpretation of Aerial Images and Maps Using a Digital Landscape Model as Partial Interpretation. In: Förstner W, Plümer L (eds) *Semantic Modeling for the Acquisition of Topografic Information from Images and Maps* (pp. 3–19). Basel: Birkhäuser.

Kohli P, Ladicky L, Torr PH (2008) Robust Higher Order Potentials for Enforcing Label Consistency. In: *IEEE Conference on Computer Vision and Pattern Recognition*

Kohli P, Ladicky L, Torr PH (2009) Robust Higher Order Potentials for Enforcing Label Consistency. *International Journal of Computer Vision*, 82 (3): 302–324.

Kolmogorov V, Zabih R (2004) What Energy Functions Can Be Minimized via Graph Cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26 (2): 147–159.

Korč F, Förstner W (2008) Interpreting terrestrial images of urban scenes using discriminative random fields. In: *International Archives of Photogrammetry and Remote Sensing*. ISPRS Symposium Beijing 2008, 37: 291–296.

Korč F, Förstner W (2009) *eTRIMS Image Database for Interpreting Images of Man-Made Scenes*. Technical Report TR-IGG-P-2009-01.

Kraus K (2007) *Photogrammetry*. Walter de Gruyter, 2nd edition.

Kschischang FR, Frey BJ, Loeliger H (2001) Factor Graphs and the Sum-Product Algorithm. *IEEE Transactions on Information Theory*, 47 (2): 498–519.

Kumar S, Hebert M (2003) Discriminative random fields: A discriminative framework for contextual interaction in classification. In: *IEEE International Conference on Computer Vision*. 2: 1150–1157.

Kumar S, Hebert M (2005) A Hierarchical Field Framework for Unified Context-Based Classification. In: *IEEE International Conference on Computer Vision*

Kumar S, Hebert M (2006) Discriminative Random Fields. *International Journal of Computer Vision*, 68 (2): 179–201.

Kunz D, Schilling KJ, Vögtle T (1997) A new approach for satellite image analysis by means of a semantic network. In: Förstner W, Plümer L (eds) *Semantic Modeling for the Acquisition of Topografic Information from Images and Maps* (pp. 20–36). Basel: Birkhäuser.

Ladicky L, Russell C, Kohli P, Torr PH (2009) Associative Hierarchical CRFs for Object Class Image Segmentation. In: *IEEE International Conference on Computer Vision*

Ladicky L, Russell C, Kohli P, Torr PH (2010) Graph Cut based Inference with Co-occurrence Statistics. In: *European Conference on Computer Vision*

Lafferty J, McCallum A, Pereira F (2001) Conditional Random Fields: Probabilistic Models for segmenting and labeling sequence data. In: *International Conference on Machine Learning*

Leberl FW (1990) *Radargrammetric image processing.* Artech House.

Li SZ (2009) *Markov Random Field Modeling in Image Analysis.* In: Sameer Singh: *Advances in Pattern Recognition.* Berlin: Springer, 3rd edition.

Lillesand TM, Kiefer RW, Chipman JW (2008) *Remote Sensing and Image Interpretation.* John Wiley and Sons, Inc, 6th edition.

Lim JJ, Arbelaez P, Gu C, Malik J (2009) Context by Region Ancestry. In: *IEEE International Conference on Computer Vision*

Liu DC, Nocedal J (1989) On the Limited Memory BFGS method for large scale optimization. *Mathematical Programming*, 45: 503–528.

Lombardo P, Oliver CJ, Macri-Pellizzeri T, Meloni M (2003) A New Maximum-Likelihood Joint Segmentation Techique for Multitemporal SAR and Multiband Optical Images. *IEEE Transactions on Geoscience and Remote Sensing*, 41 (11): 2500–2518.

Lu WL, Murphy KP, Little JJ, Sheffer A, Fu H (2009) A Hybrid Conditional Random Field for Estimating the Underlying Ground Surface from Airborne LiDAR Data. *IEEE Transactions on Geoscience and Remote Sensing*, 47 (8): 2913–2922.

Macri-Pellizzeri T, Oliver CJ, Lombardo P (2002) Segmentation-based joint classification of SAR and optical images. *IEE Radar Sonar Navigation*, 149 (6): 281–296.

Meier E, Frei U, Nüesch D (1993) *Precise Terrain Corrected Geocoded Images.* In: G. Schreier: *SAR Geocoding: Data and Systems.* Wichmann, Karlsruhe, Germany.

Michaelsen E, Stilla U (2002) Probabilistic Decisions in Production Nets: An Example from Vehicle Recognition. In: Caelli T, Amin A, Duin RP, Kamel M, de Ridder D (eds) *Advances in structural and syntactical pattern recognition, LNCS 2396* (pp. 225–233). Berlin: Springer.

Mikhail EM (1976) *Observations and least squares.* IEP Dun-Donnelley.

Morris J, Fosler-Lussier E (2008) Conditional Random Fields for Integrating Local Discriminative Classifiers. *IEEE Transactions on Audio, Speech, and Language Processing*, 16 (3): 617–628.

Murphy KP, Torralba A, Freeman WT (2004) Using the Forest to See the Trees: A Graphical Model Relating Features, Objects, and Scenes. In: Thrun S, Saul L, Schölkopf B (eds) *Advances in Neural Information Processing Systems.* Cambridge, MA: MIT Press.

Niemeier W (2002) *Ausgleichungsrechnung.* Walter de Gruyter.

Nigam K, Lafferty J, McCallum A (1999) Using maximum entropy for text classification. In: *Workshop on Machine Learning for Information Filtering*: 61–67.

Nocedal J (1980) Updating Quasi-Newton Matrices with Limited Storage. *Mathematics of Computation*, 35: 773–782.

Nocedal J, Wright SJ (2006) *Numerical Optimization.* Berlin: Springer.

Oliva A, Torralba A (2007) The role of context in object recognition. *Trends in Cognitive Sciences*, 11 (12): 520–527.

Opelt A, Pinz A, Fussenegger M, Auer P (2006) Generic Object Recognition with Boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28 (3): 416–431.

Palmer SE (1975) The effects of contextual scenes on the identification of objects. *Memory and Cognition*, 3 (5): 519–526.

Pearl J (1982) Reverend Bayes on inference engines: A distributed hierarchical approach. In: *National Conference on Artificial Intelligence*

Pearl J (1988) *Probabilistic reasoning in intelligent systems: networks of plausible inference.* San Francisco: Morgan Kaufmann, 2nd edition.

Potter MC (1975) Meaning in Visual Search. *Science*, 187 (4180): 965–966.

Poulain V, Inglada J, Spigai M (2008) High resolution remote sensing image analysis with exogenous data: a generic framework. In: *IEEE International Geoscience and Remote Sensing Symposium.* 2: 1025–1028.

Poulain V, Inglada J, Spigai M, Tourneret J, Marthon P (2009) Fusion of high resolution optical and SAR images with vector data bases for change detection. In: *IEEE International Geoscience and Remote Sensing Symposium*

Poulain V, Inglada J, Spigai M, Tourneret JY, Marthon P (2011) High-Resolution Optical and SAR Image Fusion for Building Database Updating. *IEEE Transactions on Geoscience and Remote Sensing.*

Quartulli M, Datcu M (2004) Stochastic Geometrical Modeling for Built-Up Area Understanding From a Single SAR Intensity Image With Meter Resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 42 (9): 1996–2003.

Quattoni A, Wang S, Morency LP, Collins M, Darrell T (2007) Hidden Conditional Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29 (10): 1848–1853.

Quint F, Sties M (1996) An evidential merit function to guide search in a semantic network based image analysis system. In: Perner P, Wang P, Rosenfeld A (eds) *Advances in structural and syntactical pattern recognition, LNCS 1121* (pp. 140–149). Berlin: Springer.

Rabinovich A, Vedaldi A, Galleguillos C, Wiewiora E, Belongie S (2007) Objects in Context. In: *IEEE International Conference on Computer Vision*

Raggam J, Strobl D, Hummelbrunner W (1993) *Product Quality Enhancement and Quality Evaluation.* In: G. Schreier: *SAR Geocoding: Data and Systems.* Wichmann, Karlsruhe, Germany.

Roscher R, Waske B, Förstner W (2010) Kernel Discriminative Random Fields for Land Cover Classification. In: *6th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS 2010), Istanbul, Turkey*

Rottensteiner F, Trinder J, Clode S, Kubik K (2007) Building detection by fusion of airborne laserscanner data and multi-spectral images: Performance evaluation and sensitivity analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62 (2): 135–149.

Rymes M (2000) NREL SOLPOS. `http://www.nrel.gov/midc/solpos/` accessed 10th June 2010, software developed at the National Renewable Energies Laboratory, USA.

Savarese S, Winn J, Criminisi A (2006) Discriminative Object Class Models of Appearance and Shape by Correlatons. In: *IEEE Conference on Computer Vision and Pattern Recognition*

Schistad AH, Taxt T, Jain AK (1996) A Markov Random Field Model for Classification of Multisource Satellite Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 34 (1): 100–113.

Schmitt M, Stilla U (2011) Fusion of Airborne Multi-Aspect InSAR Data by Simultaneous Backward Geocoding. In: *IEEE/GRSS/ISPRS Joint Urban Remote Sensing Event*: 53–56.

Schwäbisch M, Moreira J (1999) The High Resolution Airborne Interferometric SAR AeS-1. In: *Fourth International Airborne Remote Sensing Conference and Exhibition*: 540–547.

Serpico SB, Roli F (1995) Classification of Multisensor Remote-Sensing Images by Structured Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing*, 33 (3): 562–578.

Sha F, Pereira F (2003) Shallow parsing with conditional random fields. In: *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology.* HLT-NAACL, 1: 134–141.

Shafer G (1976) *A mathematical theory of evidence.* Princeton: Princeton University Press.

Shi J, Malik J (1997) Normalized Cuts and Image Segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition*: 731–737.

Shi J, Malik J (2000) Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 (8): 888–905.

Shotton J, Winn J, Rother C, Criminisi A (2006) TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation. In: Leonardis A, Bischof H, Pinz A (eds) *ECCV 2006, LNCS 3951* (pp. 1–15). Berlin: Springer.

Simonetto E, Oriot H, Garello R (2005) Rectangular Building Extraction From Stereoscopic Airborne Radar Images. *IEEE Transactions on Geoscience and Remote Sensing*, 43 (10): 2386–2395.

Simonetto E, Oriot H, Garello R, Le Caillec J (2003) Radargrammetric Processing for 3-D Building Extraction from High-Resolution Airborne SAR Data. In: *IEEE International Geoscience and Remote Sensing Symposium.* 3: 2002–2004.

Soergel U, ed (2010) *Radar Remote Sensing of Urban Areas.* Berlin: Springer.

Soergel U, Cadario E, Thiele A, Thoennessen U (2008) Feature Extraction and Visualization of Bridges over Water from high-resolution InSAR

Data and one Orthophoto. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1 (2): 147–153.

Soergel U, Michaelsen E, Thiele A, Cadario E, Thoennessen U (2009) Stereo analysis of high-resolution SAR images for building height estimation in cases of orthogonal aspect directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64: 490–500.

Soergel U, Thoennessen U, Stilla U (2003a) Iterative Building Reconstruction in Multi-Aspect InSAR Data. In: Maas HG, Vosselman G, Streilein A (eds) *International Archives of Photogrammetry and Remote Sensing.* 34: 186–192.

Soergel U, Thoennessen U, Stilla U (2003b) Reconstruction of Buildings from Interferometric SAR Data of built-up Areas. In: *International Archives of Photogrammetry and Remote Sensing.* 3/W8: 59–64.

Sportouche H, Tupin F, Denise L (2009) Building Extraction and 3D Reconstruction in Urban Areas from High-Resolution Optical and SAR Imagery. In: *IEEE/GRSS/ISPRS Joint Urban Remote Sensing Event*

Sportouche H, Tupin F, Denise L (2011) A symmetric scheme for building reconstruction from a couple of HR optical and SAR data. In: *IEEE/GRSS/ISPRS Joint Urban Remote Sensing Event*: 209–212.

Stilla U (1995) Map-aided structural analysis of aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 50 (4): 3–10.

Stilla U, Hedman K (2010) Feature Fusion Based on Bayesian Network Theory for Automatic Road Extraction. In: Soergel U (ed) *Radar Remote Sensing of Urban Areas* (pp. 69–86). Berlin: Springer.

Suri S, Reinartz P (2010) Mutual-Information-Based Registration of TerraSAR-X and Ikonos Imagery in Urban Areas. *IEEE Transactions on Geoscience and Remote Sensing*, 48 (2): 939–949.

Suri S, Türmer S, Reinartz P, Stilla U (2009) Registration of High Resolution SAR and Optical Satellite Imagery in Urban Areas. In: *IntArchPhRS*. 38.

Sutton C, McCallum A (2006) An Introduction to Conditionall Random Fields for Relational Learning. In: Getoor L, Taskar B (eds) *Introduction to Statistical Relational Learning.* Cambridge, MA: MIT Press.

Szeliski R, Zabih R, Scharstein D, Veksler O, Kolmogorov V, Agarwala A, Tappen M, Rother C (2008) A Comparative Study of Energy Minimization Methods for Markov Random Fields with Smoothness-Based Priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30 (6): 1068–1080.

Thiele A, Cadario E, Schulz K, Soergel U (2010a) Analysis of Gable-Roofed Building Signature in Multiaspect InSAR Data. *IEEE Geoscience and Remote Sensing Letters*, 7 (1): 83–87.

Thiele A, Cadario E, Schulz K, Thoennessen U, Soergel U (2007a) Building Recognition From Multi-Aspect High-Resolution InSAR Data in Urban Areas. *IEEE Transactions on Geoscience and Remote Sensing*, 45 (11): 3583–3593.

Thiele A, Cadario E, Schulz K, Thoennessen U, Soergel U (2007b) Feature Extraction of Gable-Roofed Buildings from Multi-Aspect High-Resolution InSAR Data. In: *IEEE International Geoscience and Remote Sensing Symposium*: 262–265.

Thiele A, Wegner JD, Soergel U (2010b) Building Reconstruction from Multi-Aspect InSAR Data. In: Soergel U (ed) *Radar Remote Sensing of Urban Areas* (pp. 187–214). Berlin: Springer.

Torralba A (2003) Contextual Priming for Object Detection. *International Journal of Computer Vision*, 53 (2): 169–191.

Torralba A, Murphy KP, Freeman WT (2005) Contextual Models for Object Detection Using Boosted Random Fields. In: Saul LK, Weiss Y, Bottou L (eds) *Advances in Neural Information Processing Systems* (pp. 1401–1408). Cambridge, MA: MIT Press.

Torralba A, Murphy KP, Freeman WT, Rubin MA (2003) Context-based vision system for place and object recognition. In: *IEEE International Conference on Computer Vision*

Toutin T (1995) Multisource data fusion with an integrated and unified geometric modelling. *EARSeL Journal: Advances in Remote Sensing*, 4 (2): 118–129.

Tóvári D, Vögtle T (2004) Classification methods for 3d objects in laserscanning data. In: *International Archives of Photogrammetry and Remote Sensing.* 35.

Tupin F (2003) Extraction of 3D information using overlay detection on SAR images. In: *2nd GRSS/ISPRS Joint Workshop on Data Fusion and Remote Sensing over Urban Areas*: 72–76.

Tupin F, Maître H, Mangin JF, Nicolas JM, Pechersky E (1998) Detection of Linear Features in

SAR Images: Application to Road Network Extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 36 (2): 434–453.

Tupin F, Roux M (2003) Detection of building outlines based on the fusion of SAR and optical features. *ISPRS Journal of Photogrammetry and Remote Sensing*, 58: 71–82.

Tupin F, Roux M (2005) Markov Random Field on Region Adjacency Graph for the Fusion of SAR and Optical Data in Radargrammetric Applications. *IEEE Transactions on Geoscience and Remote Sensing*, 43 (8): 1920–1928.

Vedaldi A, Soatto S (2008) Quick shift and kernel methods for mode seeking. In: *ECCV*

Vishwanathan S, Schraudolph NN, Schmidt MW, Murphy KP (2006a) Accelerated Training of Conditional Random Fields with Stochastic Gradient Methods. In: *International Conference on Machine Learning*: 969–976.

Vishwanathan S, Schraudolph NN, Schmidt MW, Murphy KP (2006b) CRF2D toolbox. `http://www.cs.ubc.ca/~murphyk/Software/CRF/crf2D_usage.html`.

Waske B, Benediktsson JA (2007) Fusion of Support Vector Machines for Classification of Multisensor Data. *IEEE Transactions on Geoscience and Remote Sensing*, 45 (12): 3858–3866.

Waske B, van der Linden S (2008) Classifying Multilevel Imagery From SAR and Optical Sensors by Decision Fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 46 (5): 1457–1466.

Wegner JD (2007) Automatic Fusion of SAR and Optical Imagery. Thesis in partial fullfillment of requirements for the degree of Dipl.-Ing.; accomplished at the Centre National d'Études Spatiales (CNES), Toulouse, France.

Wegner JD, Auer S, Soergel U (2010) Extraction and Geometrical Accuracy of Double-bounce Lines in High Resolution SAR Images. *Photogrammetric Engineering and Remote Sensing*, 76 (9): 1071–1080.

Wegner JD, Haensch R, Thiele A, Soergel U (2011a) Building Detection from One Orthophoto and High-Resolution InSAR Data Using Conditional Random Fields. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 4 (1): 83–91.

Wegner JD, Rosenhahn B, Soergel U (2011b) Implicit scene context for object segmentation and classification. In: *33rd Annual Symposium of the German Association for Pattern Recognition (DAGM)* accepted for publication.

Wegner JD, Rosenhahn B, Soergel U (2011c) Segment-based building detection with Conditional Random Fields. In: *6th IEEE/GRSS/ISPRS Joint Urban Remote Sensing Event*: 205–208.

Wegner JD, Thiele A, Soergel U (2009) Fusion of optical and InSAR features for building recognition in urban areas. In: *International Archives of Photogrammetry and Remote Sensing*. CMRT09, 38: 169–174.

Wolf L, Bileschi S (2006) A Critical View of Context. *International Journal of Computer Vision*, 69 (2): 251–261.

Xiao R, Lesher C, Wilson B (1998) Building Detection and Localization Using a Fusion of Interferometric Synthetic Aperture Radar and Multispectral image. In: *ARPA Image Understanding Workshop*: 583–588.

Xu F, Jin YQ (2007) Automatic Reconstruction of Building Objects From Multiaspect Meter-Resolution SAR Images. *IEEE Transactions on Geoscience and Remote Sensing*, 45 (7): 2336–2353.

Yedidia JS, Freeman WT, Weiss Y (2005) Constructing Free-Energy Approximations and Generalized Belief Propagation Algorithms. *IEEE Transactions on Information Theory*, 51 (7): 2282–2312.

Zadeh LA (1965) Fuzzy sets. *Information and Control*, 8 (3): 338–353.

Zhong P, Wang R (2007) A Multiple Conditional Random Fields Ensemble Model for Urban Area Detection in Remote Sensing Optical Images. *IEEE Transactions on Geoscience and Remote Sensing*, 45 (12): 3978–3988.

Zhu XX, Bamler R (2010) Very High Resolution Spaceborne SAR Tomography in Urban Environment. *IEEE Transactions on Geoscience and Remote Sensing*, 48 (12): 4296–4308.

Ziehn JR (2010) Accuracy Analysis of Building Height Extraction from Fusion of SAR and Optical Imagery. Bachelor Thesis.

A. Training and Inference

Random field techniques call for training and inference, which are computed iteratively. Training means the adjustment of parameters within an optimization framework, conducted based on semantically annotated data. Inference is done for computation of probabilities, mainly the partition function. It is needed twice, during training as well as to generate probabilities for testing data. The following sections give a short introduction to both steps.

A.1. Training

The objective of training is to adjust parameters of the classifier function such that classes are discriminated in an optimal way. In this thesis, object detection is viewed as a binary classification (e.g., building versus background), the task is to find an optimal decision surface in feature space separating both classes. Parameters to be trained model shape, orientation, and position of this surface. They are the elements of node weight vector $\mathbf{w}$ (Eq. 2.17) and of edge weight vector $\mathbf{v}$ (Eq. 2.18). In order to ease notation, one can concatenate parameters of $\mathbf{w}$ and $\mathbf{v}$ in a single parameter vector $\boldsymbol{\theta} = (w_1, w_2, ..., w_n; v_1, v_2, ..., v_m)$ with number of node features n and number of edge features m. Similarly, feature vectors $\mathbf{h}(\mathbf{x})$ and $\boldsymbol{\mu}(\mathbf{x})$ are concatenated to one vector $\mathbf{\Phi}$. In case of binary classification, labels -1 and 1 occur in $\mathbf{y}$. Since all features have been scaled in a range between 0 and 1, feature values $\mathbf{\Phi}$ multiplied with $\mathbf{y}$ range from -1 to 1. Each node in training data is either labeled -1 or 1 and thus all elements of the product $y_i\mathbf{\Phi}_i$ of a particular node i are either in range $[-1, 0]$ or $[0, 1]$. The general CRF as introduced in equation 2.15, where potentials are expressed with linear models (cf. Eq. 2.17 & 2.18), can then be rewritten as[1]:

$$P(\mathbf{y}|\mathbf{\Phi}, \boldsymbol{\theta}) = \frac{\exp\left(\mathbf{y}\boldsymbol{\theta}^T\mathbf{\Phi}\right)}{\sum_{\mathbf{y}} \exp\left(\mathbf{y}\boldsymbol{\theta}^T\mathbf{\Phi}\right)} \tag{A.1}$$

The denominator corresponds to the partition function (Eq. 2.16). A quotient of exponentials with linear models $\mathbf{y}\boldsymbol{\theta}^T\mathbf{\Phi}$ is the so-called *softmax function* [Bishop, 2006, chap. 4].

Adjustment of parameters $\boldsymbol{\theta}$ is an unconstrained nonlinear optimization problem that has to search a very large space of parameters. Being an entire research area of its own and since focus of

[1] Note that the posterior probability $P(\mathbf{y}|\mathbf{\Phi}, \boldsymbol{\theta})$ is abbreviated in all previous sections to $P(\mathbf{y}|\mathbf{x})$. Parameters $\boldsymbol{\theta}$ are not explicitly written and data $\mathbf{x}$ instead of concatenated features of association and interaction potential $\mathbf{\Phi}$ are set.

this thesis is on context modelling and not on designing optimization techniques, a state-of-the-art method as used in [Vishwanathan et al., 2006a,b] is applied. It couples the optimization method *Limited Memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS)* [Nocedal, 1980; Liu & Nocedal, 1989; Nocedal & Wright, 2006] with inference via *Loopy Belief Propagation (LBP)* [Frey & MacKay, 1998; Bishop, 2006] for training. This section provides an overview of this training technique. The reader is referred to literature specializing on nonlinear optimization for a more detailed and comprehensive explanation (e.g., [Nocedal & Wright, 2006; Bartholomew-Biggs, 2008; Hendrix & G.-Toth, 2010]).

In order to achieve an optimal decision surface in a Bayesian sense, *maximum a posteriori (MAP)* estimates of parameters $\boldsymbol{\theta}$ are derived. In equation A.2 a Bayes-estimator providing MAP estimates for all parameters of $\boldsymbol{\theta}$ conditioned on features $\boldsymbol{\Phi}$ and labels $\mathbf{y}$ is given.

$$P\left(\boldsymbol{\theta}|\boldsymbol{\Phi},\mathbf{y}\right) = \frac{P\left(\mathbf{y}|\boldsymbol{\Phi},\boldsymbol{\theta}\right)\cdot P\left(\boldsymbol{\Phi}|\boldsymbol{\theta}\right)\cdot P\left(\boldsymbol{\theta}\right)}{P\left(\boldsymbol{\Phi},\mathbf{y}\right)} \tag{A.2}$$

Assuming a uniform distribution of features $\boldsymbol{\Phi}$ conditioned on parameters $\boldsymbol{\theta}$ and neglecting normalization through $P\left(\boldsymbol{\Phi},\mathbf{y}\right)$, MAP estimates $P\left(\boldsymbol{\theta}|\boldsymbol{\Phi},\mathbf{y}\right)$ can be formulated as [Vishwanathan et al., 2006a]:

$$P\left(\boldsymbol{\theta}|\boldsymbol{\Phi},\mathbf{y}\right) \propto P\left(\mathbf{y}|\boldsymbol{\Phi},\boldsymbol{\theta}\right)\cdot P\left(\boldsymbol{\theta}\right) \tag{A.3}$$

$P\left(\mathbf{y}|\boldsymbol{\Phi},\boldsymbol{\theta}\right)$ is the posterior probability of the CRF as given in equation A.1 and $P\left(\boldsymbol{\theta}\right)$ is a prior over parameters $\boldsymbol{\theta}$ acting as a regularization term. It penalizes large parameters thus smoothing the objective function to avoid over-fitting to training data. Usually, the assumption is made that parameters $\boldsymbol{\theta}$ follow an isotropic Gaussian prior[2] [Vishwanathan et al., 2006a; Sutton & McCallum, 2006] and one may thus write their probability as $P\left(\boldsymbol{\theta}\right) = \exp\left(\frac{-(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^2}{2\sigma^2}\right)$ with $\boldsymbol{\theta}_0 = 0$. This leads to a regularization term containing euclidean norm $\|\boldsymbol{\theta}\|$ of parameters and variance σ^2:

$$P\left(\boldsymbol{\theta}\right) = \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{\theta}\|^2\right) \tag{A.4}$$

The choice of σ steers smoothness of the objective function with respect to training data. A larger σ results in a smoother function whereas a smaller σ better adapts the objective function to training data, but at the risk of over-fitting[3]. An appropriate objective function, ensuring exactly one global optimum, has to be designed to obtain $P\left(\boldsymbol{\theta}|\boldsymbol{\Phi},\mathbf{y}\right)$. It should either be convex (global minimum) or concave (global maximum). A concave objective function can be reformulated as a convex function and vice versa. This criterion is met using the *regularized log likelihood* as objective function, like

[2] In literature (e.g., [Vishwanathan et al., 2006a]) this Gaussian prior is motivated via computational convenience, a theoretical justification is missing. Investigation of this issue is left for future work.

[3] Tests with different values σ between one and 100 did not lead to significant changes of classification results. This effect is also observed by Sutton & McCallum [2006, chap. 1.3.2]. One possible explanation is that the problem of the CRF in its present form is not over-fitting, but under-fitting. Classification results (cf. section 4.2) hint in this direction, too, unveiling problems with over-smoothing (i.e., a too smooth decision surface in feature space). Smoothing does not have essential impact on classification because the decision surface already is too smooth.

done by Lafferty et al. [2001], Kumar & Hebert [2003] and Sutton & McCallum [2006]. Objective function $L(\boldsymbol{\theta})$ to be optimized for MAP parameter estimation is:

$$L(\boldsymbol{\theta}) = \log\left(P\left(\mathbf{y}|\boldsymbol{\Phi},\boldsymbol{\theta}\right)\cdot P\left(\boldsymbol{\theta}\right)\right) \tag{A.5}$$

Substituting the right side of equation A.1 for $P(\mathbf{y}|\boldsymbol{\Phi},\boldsymbol{\theta})$ and the right side of equation A.4 for $P(\boldsymbol{\theta})$, the penalized log likelihood function is:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \log\left(\frac{\exp\left(\mathbf{y}\boldsymbol{\theta}^T\boldsymbol{\Phi}\right)}{\sum_{\mathbf{y}}\exp\left(\mathbf{y}\boldsymbol{\theta}^T\boldsymbol{\Phi}\right)}\cdot\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{\theta}\|^2\right)\right) \\ &= \mathbf{y}\boldsymbol{\theta}^T\boldsymbol{\Phi} - \log\sum_{\mathbf{y}}\exp\left(\mathbf{y}\boldsymbol{\theta}^T\boldsymbol{\Phi}\right) - \frac{1}{2\sigma^2}\|\boldsymbol{\theta}\|^2 \end{aligned} \tag{A.6}$$

The global maximum of the objective function in equation A.6 has to be found in order to achieve MAP estimates of parameters $\boldsymbol{\theta}$. Alternatively, the global minimum of the *negative penalized log likelihood function* (taking the negative logarithm in Eq. A.5) can be sought with an appropriate optimization technique.

In general, approaches making use of the second derivative (i.e., curvature of the objective function) like the Newton method converge reasonably fast on large scale optimization tasks. Computing second derivatives implies an analytical expression of second partial derivatives of all parameters to set up the Hessian matrix at each iteration. It becomes unpractical if dealing with a huge number of parameters [Sutton & McCallum, 2006, chap. 1.3.2] as in this case[4].

A common way to gain faster convergence using curvature information, but circumventing the need for explicitly computing a Hessian, are so-called *Quasi-Newton methods*. They approximate second derivatives exploiting the iterative update scheme of first derivatives of the objective function. The Limited Memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method [Nocedal, 1980; Liu & Nocedal, 1989; Nocedal & Wright, 2006], used for optimization in this thesis, is such a Quasi-Newton method approximating second derivatives via differences of first derivatives. It has shown to provide good results for CRFs (e.g., [Sha & Pereira, 2003; Vishwanathan et al., 2006a; Morris & Fosler-Lussier, 2008]).

Instead of analytically deriving first partial derivatives of the objective function (Eq. A.6), gradients are approximated via differences between expected feature vectors of cliques $\mathbb{E}(\mathbf{y}\boldsymbol{\Phi})$ and original feature vectors $\mathbf{y}\boldsymbol{\Phi}$ at a training iteration. In case of the linear model used here, all feature vector elements are multiplied with binary labels $(-1, 1)$.

The first derivative of the objective function (Eq. A.6) with respect to parameters $\boldsymbol{\theta}$ has to be computed at each training iteration as input to the optimizer. In case of $\mathbf{y}\boldsymbol{\theta}^T\boldsymbol{\Phi}$ we get $\mathbf{y}\boldsymbol{\Phi}$

[4] It should be reconsidered that each feature is assigned a weight. Increasing the number of features leads to an increasing number of parameters to be learned, too.

(the original feature vector) and the first derivative of the Gaussian prior is $\boldsymbol{\theta}/\sigma^2$. First partial derivation of the middle term of equation A.6, the logarithm of partition function $\log Z(\mathbf{\Phi}) = \log \sum_{\mathbf{y}} \exp\left(\mathbf{y}\boldsymbol{\theta}^T\mathbf{\Phi}\right)$, leads to expected feature vectors $\mathbb{E}(\mathbf{y}\mathbf{\Phi})$:

$$\begin{aligned}\frac{\partial Z}{\partial \boldsymbol{\theta}} &= \frac{1}{\sum_{\mathbf{y}} \exp\left(\mathbf{y}\boldsymbol{\theta}^T\mathbf{\Phi}\right)} \cdot \sum_{\mathbf{y}} \exp\left(\mathbf{y}\boldsymbol{\theta}^T\mathbf{\Phi}\right) \cdot \mathbf{y}\mathbf{\Phi} \\ &= \sum_{\mathbf{y}} \frac{\exp\left(\mathbf{y}\boldsymbol{\theta}^T\mathbf{\Phi}\right) \cdot \mathbf{y}\mathbf{\Phi}}{\sum_{\mathbf{y}} \exp\left(\mathbf{y}\boldsymbol{\theta}^T\mathbf{\Phi}\right)} \\ &= \sum_{\mathbf{y}} P(\mathbf{y}|\boldsymbol{\theta},\mathbf{\Phi}) \cdot \mathbf{y}\mathbf{\Phi} \\ &= \mathbb{E}(\mathbf{y}\mathbf{\Phi})\end{aligned} \tag{A.7}$$

The first factor in the top row of equation A.7 stems from taking the first derivative of the logarithm, the second one is the derivative of the exponential function, and the third one is the inner derivation of the exponent of the exponential function. Expressing the first row in a more compact form results in the second row. It reveals that the quotient $\exp\left(\mathbf{y}\boldsymbol{\theta}^T\mathbf{\Phi}\right) / \sum_{\mathbf{y}} \exp\left(\mathbf{y}\boldsymbol{\theta}^T\mathbf{\Phi}\right)$ is exactly the posterior probability $P(\mathbf{y}|\boldsymbol{\theta},\mathbf{\Phi})$ of the CRF as given in equation A.1. Substituting the posterior probability leads to the third row. Summing up all possible products $P(\mathbf{y}|\boldsymbol{\theta},\mathbf{\Phi}) \cdot \mathbf{y}\mathbf{\Phi}$ for all values of $\mathbf{y}$ is the expected feature vector (multiplied with labels) $\mathbb{E}(\mathbf{y}\mathbf{\Phi})$ of a clique.

The first partial derivative of the objective function given in the second row of equation A.6 is:

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = \mathbf{y}\mathbf{\Phi} - \mathbb{E}(\mathbf{y}\mathbf{\Phi}) - \frac{\boldsymbol{\theta}}{\sigma^2} \tag{A.8}$$

Posterior probabilities $P(\mathbf{y}|\boldsymbol{\theta},\mathbf{\Phi})$ of labels conditioned on parameters and features are computed with LBP inference. It is this training step that intertwins inference with optimization in order to determine the first partial derivatives of the objective function. LBP inference is only necessary to compute the expected feature vectors as shown in the third row of equation A.7. The first partial derivatives are needed as input to optimization with L-BFGS. In addition to the gradient, L-BFGS needs the function value of the objective function (Eq. A.6) at each training iteration. First and third term in equation A.6 can directly be computed, but this is not the case with the middle term, the partition function.

Partition function $Z(\mathbf{\Phi})$ sums over all possible label configurations $\mathbf{y}$ of features $\mathbf{\Phi}$. In general, it is computationally very costly to calculate the global partition function directly due to the very high number of possible label configurations $\mathbf{y}$. In addition, it has to be evaluated anew at each training iteration. In case of binary classification, for example, a node can be labeled with two different labels. If we have two nodes in the graph and labels 1 and -1, four different configurations of labels exist ([1,1], [-1,-1], [-1,1], [1,-1]). One additional node in the graph leads to eight different configurations. For a binary labeling task with n nodes in the graph we obtain 2^n possible label configurations.

Considering that we have 2500 nodes if subdividing one of the 1000x1000 pixels orthophoto test images (cf. Fig. 4.3(a)-(c)) into 20×20 pixels patches, we obtain 2^{2500} possible label configurations. Evaluating $\log Z(\mathbf{\Phi})$ for 2^n possible label configurations at each training iteration becomes infeasible and thus it is usually approximated. One possibility to approximate $\log Z(\mathbf{\Phi})$ is to use Bethe Free Energy (BFE) approximations originating in physics. The reader is referred to [Yedidia et al., 2005] and related literature for a description of BFE approximation. Function value $\log Z(\mathbf{\Phi})$ is evaluated via BFE at all training iterations and combined with Loopy Belief Propagation that passes on the belief values after inference at the current training step.

Having computed, both, gradients and function values of the objective function (Eq. A.6), inputs to L-BFGS optimization are complete. They have to be recomputed for every training iteration. A detailed description of the L-BFGS algorithm can be found in chapter 6 of [Nocedal & Wright, 2006].

A.2. Inference

In general, probabilistic inference determines the probability that a hypothesis may be true given some observations. Inference is necessary in graphical models to compute marginal probabilities for all nodes in the graph. Reconsidering directed, undirected, and factor graph given in figure 2.9, one probability per node and class has to be determined. Dealing with binary classification, for example, probabilistic inference computes two marginal probabilities at every node, one for the first and one for the second class. A MAP estimator would then assign the class with highest marginal probability to a node. Inference is needed twice in CRFs: first, during training for gradient computation (cf. Eq. A.7) and, second, to find the (posterior) marginal probability of each class during testing.

Various approaches to probabilistic inference exist, many relying on message passing in graphs to compute marginals. In this thesis, so-called *Loopy Belief Propagation (LBP)* is used. It basically applies Belief Propagation, originally developed for inference in graphs with tree-like structures [Pearl, 1982, 1988], to undirected graphs with cycles like the CRF. LBP is a message passing algorithm minimizing the energy within a graph by passing messages from nodes via edges to neighbouring nodes. Moreover, it is a particular form of the sum-product algorithm [Bishop, 2006, chap. 8.4.4], which is used for exact inference in trees.

One important drawback of LBP is that it may end up in a local extremum, global convergence is not guaranteed. Nonetheless, it is a widely used standard technique and Frey & MacKay [1998] showed that although originally developed for trees, LBP usually well approximates the global optimum (a theoretical explanation is given by Yedidia et al. [2005]). [Szeliski et al., 2008] compared different state-of-the-art methods for energy minimization within Markov Random Fields and LBP was one of the best performing methods. It was chosen for all presented results in this thesis and will be briefly described in the following. Another promising approach would be graph cuts (Boykov et al. [2001]; Kolmogorov & Zabih [2004]). It is left for future work.

LBP can best be explained with a graph structured as a so-called *factor graph* [Kschischang et al., 2001] (cf. section 2.2.2). It consists of variable nodes and factor nodes (Fig. A.1(a)). The goal of inference is to label variable nodes (i.e., marginal probabilities have to be assigned), which represent spatial units like pixel, patch, or region in a graph, for example. Factor nodes are located on edges between neighbouring nodes. During inference, messages are initially sent from variable nodes to factor nodes. Then, factor nodes pass on these messages to the neighbouring nodes along edges. New values at variable and factor nodes are computed via products and sums of incoming messages. This update scheme is repeated iteratively until convergence, reached if changes of marginals at nodes are below a threshold (i.e., the *total energy within the graph* has been minimized).

This concept applies well to CRFs with association and interaction potentials as used in this thesis (cf. Eq. 2.15). The association potential acts as the initial value at a variable node, whereas the interaction potential of two neighbouring nodes is assigned to the factor node on the edge between them. In case of linear models chosen for association potential (Eq. 2.17) and interaction potential (Eq. 2.18), potentials are scalars. Considering equation 2.17, the initial node potential of variable node a in figure A.1 is $n_{pot}(a) = \exp\left(y_a \mathbf{w}^T \mathbf{h}_a(\mathbf{x})\right)$. Feature vector $\mathbf{h}_a(\mathbf{x})$ multiplied with weight vector $\mathbf{w}^T$ delivers a scalar value. It is then multiplied with a label y_a. Similarly, edge potentials are computed for factor nodes. Considering factor node f_{ab} relating nodes a and b, the interaction potential results in a scalar value $e_{pot}(a, b) = \exp\left(y_a y_b \mathbf{v}^T \boldsymbol{\mu}_{ab}(\mathbf{x})\right)$. After initial assignment of node potentials to variable nodes and edge potentials to factor nodes, messaging passing begins. A detailed description of message passing rules and update scheme are provided by Kschischang et al. [2001], Yedidia et al. [2005], and Bishop [2006, chap. 8.4.4], for example.

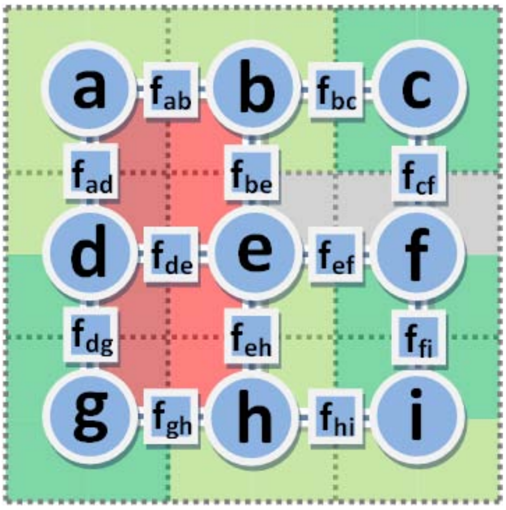


Figure A.1.: Factor graph; dotted lines represent the spatial extent of each node, nodes are shown as blue circles with white boundaries, factor nodes as squares, edges linking the nodes are represented as blue lines with white boundaries.

B. Measured and adjusted building heights

In this annex, all single measured building heights as well as the adjusted ones are provided of, both, flat roof and gable roof buildings. Numbers of flat roof buildings correspond to the ones in figure 4.14(b), gable roof building numbers are shown in figures 4.19(f) & 4.18(d,h,l).

$B\#$	h_s	h_{pd}	h_{db}	h_{InSAR}	h_l	$h_{b,noI}$	h_b	h_L	$\hat{\sigma}_b$	$\Delta_{b,L}$
1	18.7	16.0	15.7	15.1	15.1	16.9	**15.9**	**14.3**	0.72	1.6
2	9.1	-	-	-	-	-	-	**17.2**	-	-
3	22.8	21.6	22.2	23.2	23.8	23.2	**23.2**	**22.9**	0.35	0.3
4	12.7	14.9	14.4	12.5	11.8	12.4	**12.5**	**12.4**	0.25	0.1
5	17.0	15.4	19.3	15.1	17.1	17.1	**16.0**	**18.2**	0.51	-2.2
6	13.7	13.1	7.3	14.1	11.9	12.7	**13.5**	**13.9**	0.56	-0.4
7	11.6	11.0	11.3	12.4	13.2	12.2	**12.3**	**12.7**	0.28	-0.4
8	12.9	9.2	17.0	12.0	11.1	12.1	**12.0**	**12.7**	0.41	-0.7
9	9.3	11.7	8.9	10.5	8.9	9.2	**10.0**	**9.6**	0.36	0.4
10	7.9	-	9.3	10.3	9.4	8.5	**9.5**	**10.2**	0.59	-0.7
11	11.8	-	-	-	10.3	11.2	**11.2**	**12.0**	0.72	-0.8
12	3.7	-	-	-	-	-	-	**4.0**	-	-

Table B.1.: Building heights of flat roof buildings with number $B\#$ (all values in unit meter) via: sun shadow (h_s), optical perspective distortion (h_{pd}), overlap of roof edge and double-bounce line (h_{db}), robust maximum InSAR heights in layover ramp (h_{InSAR}), layover in SAR magnitude image (h_l), all single heights except InSAR heights combined with least squares adjustment ($h_{b,noI}$), all heights including InSAR heights adjusted (h_b), LiDAR reference (h_L), posterior standard deviation after adjustment ($\hat{\sigma}_b$), difference of adjusted height (including InSAR measurements) to LiDAR reference height ($\Delta_{b,L}$); note: absent height values indicate that no measurements could be conducted due to missing observations; adjusted heights are not reported if only one height could be measured at a building.

$B\#$	h_s	h_{pd}	h_r	h_b	h_L	$\hat{\sigma}_b$	$\Delta_{b,L}$
1	10.1	6.2	8.2	**9.7**	**9.7**	0.71	0
2	-	7.4	8.8	**8.4**	**8.6**	0.62	-0.2
3	-	5.6	7.8	**7.3**	**9.3**	1.00	-2.0
4	-	4.8	8.1	**7.1**	**9.1**	1.49	-2.0
5	9.9	7.5	7.4	**9.3**	**8.7**	0.72	0.6
6	9.9	6.0	8.7	**9.5**	**9.4**	0.67	0.1
7	10.1	4.9	8.9	**9.7**	**9.7**	0.85	0
8	9.3	6.1	7.9	**9.0**	**8.9**	0.57	0.1
9	-	9.7	7.6	**8.2**	**9.9**	0.93	-1.7
10	9.1	5.0	7.1	**8.5**	**9.6**	0.81	-1.1
11	9.4	7.1	8.7	**9.1**	**10.2**	0.41	-1.1
12	8.3	6.6	7.8	**8.1**	**9.3**	0.30	-1.2
20	-	-	9.3	-	**9.7**	-	-
21	-	-	9.8	-	**11.4**	-	-
22	-	-	8.9	-	**10.3**	-	-
23	-	-	7.8	-	**9.5**	-	-
24	12.9	11.6	10.3	**12.6**	**12.6**	0.58	0
25	-	-	12.2	-	**12.9**	-	-
26	-	-	11.5	-	**11.7**	-	-
27	-	-	11.8	-	**12.0**	-	-
28	-	-	11.1	-	**11.4**	-	-
29	-	-	12.4	-	**12.8**	-	-
30	13.2	9.5	11.2	**12.9**	**12.4**	0.60	0.5
31	-	-	12.0	-	**11.2**	-	-
33	-	-	12.0	-	**12.0**	-	-
34	12.1	-	11.9	**12.0**	**12.2**	0.18	-0.2
35	-	-	12.7	-	**12.6**	-	-

Table B.2.: Building heights of gable roof buildings with number $B\#$ (all values in unit meter) via: sun shadow (h_s), optical perspective distortion (h_{pd}), parallel SAR lines (h_r), all possibilities combined with least squares adjustment (h_b), LIDAR reference heights (h_L), posterior standard deviation after adjustment ($\hat{\sigma}_b$), difference of adjusted height to LiDAR reference height ($\Delta_{b,L}$); note: absent height values indicate that no measurements could be conducted due to missing observations; adjusted heights are not reported if only one height could be measured at a building.

Index

Acknowledgements

First and foremost I thank my advisor Uwe Sörgel for providing careful guidance throughout my PhD. It is due to his expertise, patience, and faith that I felt all the freedom to develop my own research ideas.

I would also like to express my sincere gratitude to Uwe Stilla and Monika Sester for acting as referees of this thesis. Despite heavy workload they agreed to review my work within a very short period of time.

Additionally, I largely benefitted from inspiring discussions with Bodo Rosenhahn concerning random fields and the implicit scene context idea, in particular. Many insights into computer vision have been initiated by him.

Moreover, I thank all my colleagues and friends at the Institute of Photogrammetry and GeoInformation (IPI) at the Leibniz Universität Hannover for their companionship. My time at IPI would not have been as inspiring without lively discussions and much appreciated critical questions. I am particularly thankful to Alexander Schunert for a great and most memorable time in our office. During the final writing phase he persuaded me to take coffee breaks when I most needed them.

I also thank Jens Ziehn who has helped me a lot with programming the height estimation framework and with processing the data. Especially in the final weeks before submission he responded quickly to any requests from my side at often unusual time of day.

Finally, I am gratefully indebted to my parents Heike and Helmut as well as to my brothers Philipp and Moritz for their absolute love, faith, and unlimited support.

CURRICULUM VITAE

PERSONAL INFORMATION

Name	**WEGNER, JAN DIRK**
Date of birth	9TH FEBRUARY 1982 IN OLDENBURG (OLDB), GERMANY

WORK EXPERIENCE

• Dates (from – to)	**SINCE 1ST OCTOBER 2007**
• Name and address of employer	Institute of Photogrammetry and GeoInformation (IPI), Leibniz Universität Hannover, Nienburger Str.1, 30167 Hannover, Germany
• Type of business or sector	Computer Vision, Photogrammetry, Remote Sensing
• Occupation or position held	Scientific collaborator
• Main activities and responsibilities	Lectures (as substitute for Uwe Sörgel) and labs of courses "Digital Image Processing" and "Radar Remote Sensing", main project work: Fusion of high-resolution Synthetic Aperture Radar (SAR) and optical data in urban areas for object detection and 3D-modeling, contextual object detection, contextual learning, graphical models
• Dates (from – to)	**5TH NOVEMBER 2006 - 4TH SEPTEMBER 2007**
• Name and address of employer	Centre National d'Études Spatiales (CNES), DCT/SI/SAR, BPI 601, 18 Avenue Edouard Belin, 31401 Toulouse Cedex 9, France
• Type of business or sector	Computer Vision, Photogrammetry, Remote Sensing
• Occupation or position held	Intern, project for graduate degree Dipl.-Ing.
• Main activities and responsibilities	Development and implementation of new methods for the open soure library Orfeo Toolbox with particular focus on automatic co-registration of high-resolution SAR and optical data in urban areas
• Dates (from – to)	**10TH JUNE 2002 - 30TH AUGUST 2002**
• Name and address of employer	Washington State Department of Transportation, Geographic Services, 1655 South 2nd Avenue, Tumwater, WA 98512, USA
• Type of business or sector	Land surveying, leveling, GPS network
• Occupation or position held	Intern
• Main activities and responsibilities	GPS, leveling and gravity measurements to link ellipsoidal heights to physical geoid heights

EDUCATION AND TRAINING

• Dates (from – to)	**1ST OCTOBER 2007 - 2ND AUGUST 2011 PHD STUDENT**
• Name and type of organisation	Fakultät für Bauingenieurwesen und Geodäsie, Leibniz Universität Hannover, Germany
• Principal subjects covered	Computer vision, photogrammetry, remote sensing, stochastics
• Title of qualification awarded	Doktor-Ingenieur (Dr.-Ing.) with distinction (summa cum laude)
• Dates (from – to)	**OCTOBER 2002 - SEPTEMBER 2007 COURSE GEODESY AND GEOINFORMATICS**
• Name and type of organisation	Leibniz Universität Hannover, Germany
• Principal subjects covered	Signal processing, stochastics, photogrammetry, remote sensing, computer vision
• Title of qualification awarded	Diplom-Ingenieur (Dipl.-Ing.), prize for best diploma thesis in 2007 awarded by Geoinformatics North (GiN e.V.)
• Dates (from – to)	**AUGUST 2001 - MAY 2002 CIVIL SERVICE**
• Name and type of organisation	Civil service with emergency medical services of Johanniter-Unfall-Hilfe in Oldenburg, Germany
• Principal subjects covered	Emergency medical aid, education as a licensed paramedic
• Dates (from – to)	**1988 -2001 HIGH SCHOOL**
• Name and type of organisation	Gundschule Bloherfelde, Orientierungsstufe Eversten, Gymnasium Eversten in Oldenburg