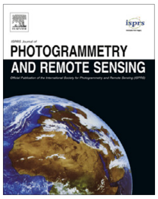
ISPRS Journal of Photogrammetry and Remote Sensing 93 (2014) 256–271

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs



Results of the ISPRS benchmark on urban object detection and 3D building reconstruction

Franz Rottensteiner [a,*], Gunho Sohn [b], Markus Gerke [c], Jan Dirk Wegner [d], Uwe Breitkopf [a], Jaewook Jung [b]

[a] Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Nienburger Straße 1, 30167 Hannover, Germany
[b] GeoICT Lab, Earth and Space Science and Engineering Department, York University, 4700 Keele St., Toronto M3J 1P3, Canada
[c] Faculty ITC, EOS Department, University of Twente, PO Box 217, 7500AE Enschede, The Netherlands
[d] Institute of Geodesy and Photogrammetry, Swiss Federal Institute of Technology Zurich, Wolfgang-Pauli-Strasse 15, 8093 Zurich, Switzerland

ARTICLE INFO

Article history:
Available online 16 November 2013

Keywords:
Automatic object extraction
3D building reconstruction
Aerial imagery
Laser scanning
Evaluation
Benchmarking test

ABSTRACT

For more than two decades, many efforts have been made to develop methods for extracting urban objects from data acquired by airborne sensors. In order to make the results of such algorithms more comparable, benchmarking data sets are of paramount importance. Such a data set, consisting of airborne image and laserscanner data, has been made available to the scientific community by ISPRS WGIII/4. Researchers were encouraged to submit their results of urban object detection and 3D building reconstruction, which were evaluated based on reference data. This paper presents the outcomes of the evaluation for building detection, tree detection, and 3D building reconstruction. The results achieved by different methods are compared and analysed to identify promising strategies for automatic urban object extraction from current airborne sensor data, but also common problems of state-of-the-art methods.

1. Introduction

The automated extraction of urban objects from data acquired by airborne sensors has been an important topic of research in photogrammetry for more than two decades (Mayer, 2008). Urban object extraction is still an active field of research, with the focus shifting to detailed representations of objects, to using data from new sensors, or to advanced processing techniques. However, scanning the relevant literature in photogrammetry and remote sensing (Schindler et al., 2011; Sohn et al., 2013; Stilla et al., 2011), it has become obvious that there is a lack of publicly available benchmark data sets with ground truth that can be used for the evaluation of their methods by the authors of research papers. As a consequence, the authors usually evaluate their methods on different data sets and using different evaluation criteria, which makes a comparison of the methods difficult and hampers a critical assessment of the pros and cons of each of the methods. In computer vision, the success of the Middlebury Stereo Vision test (Scharstein and Szeliski, 2002) and other benchmarks such as the Pascal VOC data set (Everingham et al., 2010) has shown the importance of providing common data sets with ground truth for comparing different approaches to problems such as image matching and object detection. Apart from making different approaches comparable, benchmarks can trigger progress by giving indications about the most promising strategies for the solution of a given task and by identifying common problems of existing approaches, thus showing new directions of research.

However, using standard benchmarks for object extraction from computer vision such as the Pascal VOC data set for a comparison of object extraction techniques from remote sensing imagery is not necessarily fair to the latter. Methods tailored for remote sensing data, usually characterised by vertical viewing directions, cannot rely on the availability of a reference direction such as the vertical in terrestrial images with horizontal viewing directions. Thus, on the one hand, they may perform poorly in comparison to methods that are tailored to such data and, for instance, employ priors for the location of objects in an image, modelling that the sky is usually the highest object in a scene and that roads are most likely at the bottom of an image (Yang and Förstner, 2011), whereas on the other hand, they cannot exploit features such as the NDVI that can be extracted from images taken by modern multispectral sensors. There is an obvious need for benchmark data sets consisting of airborne data that can serve as test beds for developments in the field of topographic object detection and 3D reconstruction, in particular in urban areas.

There have been attempts in the past to distribute benchmark data sets for object extraction. The authors particularly acknowledge the efforts of OEEPE/EuroSDR (European Spatial Data Research), who provided data sets for building (Kaartinen et al., 2005) and road extraction (Mayer et al., 2006) and for automated

* Corresponding author.
E-mail addresses: rottensteiner@ipi.uni-hannover.de (F. Rottensteiner), gsohn@yorku.ca (G. Sohn), m.gerke@utwente.nl (M. Gerke), jan.wegner@geod.baug.ethz.ch (J.D. Wegner), breitkopf@ipi.uni-hannover.de (U. Breitkopf), jwjung@yorku.ca (J. Jung).

0924-2716/$ - see front matter © 2013 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.isprsjprs.2013.10.004

updating of maps (Champion et al., 2009). As far as data from aerial sensors are concerned, these data sets are outdated. The first two data sets (building and road extraction) rely on scanned image; the data set for automated map updating does include one orthophoto from a modern multi-spectral camera (Champion et al., 2009), but the airborne laserscanner (ALS) data only consist of first echoes that are augmented by a scanned aerial image. There is a need for new standard test sites for urban object extraction making use of the benefits of modern airborne sensors such as multiple-overlap geometry, increased spectral and radiometric resolution of images and, in case of ALS data, the recording of multiple echoes.

These considerations led to the establishment of a benchmark on urban object extraction. A modern data set consisting of digital aerial image and ALS data along with reference data was generated and made available to the research community via the ISPRS web site (ISPRS, 2013). Unlike previous benchmark data sets on urban object detection, the reference data include 2D outlines of multiple object types and 3D roof landscapes. It also contains different types of urban development. Researchers are given access to the sensor data and encouraged to carry out one or more of several urban object extraction tasks:

Task 1: Urban object detection, i.e. the determination of the 2D outlines of urban objects in the input data. The focus of the evaluation is on the thematic and geometrical accuracy of the results.

Task 2: 3D building reconstruction, i.e. the reconstruction of detailed 3D roof structures in the test areas. The focus of evaluation is on the quality of the roof plane segmentation and on the geometrical accuracy of the roof polygons.

Up to the time of writing (October 2013), we have received results from more than 20 research groups. Some groups have submitted results achieved by different algorithms, mostly for the detection of buildings and trees and for 3D building reconstruction. It is the main goal of this paper to give an overview about the benchmark and to describe and analyse the results submitted so far, in order to highlight the potential of current state-of-the-art methods for solving the tasks. We also aim to show common problems and limitations of current methods in order to identify unsolved problems in urban object extraction. As far as object detection is concerned, we focus on buildings and trees, because hardly any results for other object types have been received so far.

This paper is structured as follows. In Section 2, we briefly present the test data sets. The tasks to be solved by the participants and the evaluation methodology are explained in Section 3. Section 4 gives a very brief overview over the methods for which we have received results. These results are evaluated and analysed in Section 5. Section 6 presents our conclusions and our plans for future work.

2. Data sets

2.1. Data set 1: Vaihingen (Germany)

This is a subset of the data used for the test of digital aerial cameras by the German Association of Photogrammetry, Remote Sensing, and Geoinformation (DGPF; Cramer, 2010). It contains twenty 16 bit pan-sharpened colour infrared (CIR) images with a ground sampling distance (GSD) of 8 cm. The images were acquired with 65% forward lap and 60% side lap using an Intergraph/ZI DMC having a focal length of 120 mm. Orientation data with pixel-level accuracy are distributed with the images. Furthermore, the data set contains ALS data acquired using a Leica ALS50 system with a point density between 4 and 7 points/m^2. Multiple pulses were recorded. A digital surface model (DSM) with a grid width of 25 cm was interpolated from the ALS points corresponding to the last echo of each pulse. Three test sites were selected for generating reference data:

Area 1 (37 buildings and 105 trees; 125 m × 200 m) is characterised by dense development consisting of historic buildings having rather complex shapes along with roads and trees.

Area 2 (14 buildings and 162 trees; 170 m × 190 m) is characterised by a few high-rising residential buildings that are surrounded by trees.

Area 3 (56 buildings and 155 trees; 150 m × 220 m) is a purely residential area with detached houses and many surrounding trees.

In these test areas, reference data were generated by manual stereo plotting. The reference for building detection consists of roof outline polygons with a planimetric accuracy of about 10 cm. The reference for tree detection consists of circles approximating the crown outlines of all trees higher than 1.5 m. The circle centres give approximate positions of the tree stems with a geometrical accuracy of about 0.5–1.5 m. The reference for building reconstruction consists of 3D building models corresponding to the level of detail LoD2 according to the CityGML standard (Gröger et al., 2008). They are detailed roof models without roof overhangs or façade details. The accuracy is about 10 cm in planimetry and height.

2.2. Data set 2: Toronto (Canada)

This second data set was captured over downtown Toronto (Canada). It also consists of image and ALS data. There are 13 RGB colour images with a radiometric resolution of 8 bit and a GSD of 15 cm. The images were acquired using a Microsoft Vexcel UltraCam-D with a focal length of 101.4 mm, arranged in a block of three strips with 60% forward lap and 30% side lap. Only stereo overlap is available in most areas. Orientation data with pixel-level accuracy are also distributed with the images. Optech's ALTM-ORION M was used to acquire the ALS data at a flying height of 650 m in 6 strips with a point density of about 6 points/m^2. A DSM with a grid width of 25 cm was interpolated from the ALS points corresponding to the last echo of each pulse. One disadvantage of this dataset is that ALS and images were not only captured in different vegetation periods (ALS: leave-on, images: leave-off), but also with a time gap of two years. Two test sites were selected in this data set:

- *Area 4* (58 buildings; 530 m × 600 m) contains a mixture of low and high-storey buildings, showing various degrees of shape complexity in rooftop structure. The scene also contains trees and other urban objects.
- *Area 5* (38 buildings; 530 m × 600 m) represents a cluster of high-rise buildings typical for the Central Business District (CBD) of North American cities. It contains shadows cast by high buildings and various types of urban objects.

The reference for building detection and for 3D building reconstruction (LoD2) was generated by stereo plotting. The accuracy of well-defined points is 20 cm in planimetry and 15 cm in height. For more details refer to (Rottensteiner et al., 2012) and to the web site of the test (ISPRS, 2013).

3. Test setup

3.1. Task 1: urban object detection

The goal of the first task is the detection of objects in the test areas. The participants can deliver outline polygons of the objects or binary object masks. For the evaluation of the thematic accuracy, the method described in (Rutzinger et al., 2009) was used. After a topological clarification required to be able to deal with the fact that in densely built-up areas building detection algorithms tend to deliver whole blocks as individual buildings, we determine the following indices to measure the quality of the results:

- $Comp_{ar}$, $Corr_{ar}$, Q_{ar}: *completeness*, *correctness* and *quality* determined on a per-area level. These indices are related to the area that was correctly classified.
- $Comp_{obj}$, $Corr_{obj}$, Q_{obj}: *completeness*, *correctness* and *quality* determined on a per-object level. These indices count the number of objects that are correctly detected. A minimum overlap of 50% for an extracted object with the reference is required for the object to be counted as a true positive. Having to choose such a threshold adds a certain degree of subjectivity to the evaluation; however, we think that an evaluation on a per-object level is important because it is more directly related to the amount of work required for post-editing the results. For a detailed discussion of this topic and also for a comparison of different thresholds, see (Rutzinger et al., 2009).
- $Comp_{50}$, $Corr_{50}$ and Q_{50}: *completeness*, *correctness* and *quality* determined on a per-object level, but only considering objects larger than 50 m^2. These indices are useful to analyse the dependency of per-object quality metrics on the object size. The threshold was chosen to select the most representative buildings per plot and the largest trees in the scene (Rottensteiner et al., 2012).
- *RMS*: The RMS error is used to characterise the geometrical quality of the detected objects. For buildings, RMS is the RMS error of the planimetric distances of the reference boundary points to their nearest neighbours on the corresponding extracted boundaries; for trees, it is the RMS error of the planimetric distances between the centres of gravity of corresponding objects in the detection results and in the reference. In both cases, only distances shorter than 3 m are considered.

3.2. Task 2: 3D building reconstruction

The goal of the second task is the generation of detailed (LoD2) 3D models of the building roofs in the test areas. The results should be submitted as closed 3D roof polygons. The evaluation focuses on an analysis of the segmentation quality and on the geometrical errors of the submitted models. In the literature, there is much less work on a quality analysis of 3D building models. Meidow and Schuster (2005) present an evaluation scheme that is an expansion of the pixel-based analysis in 2D to 3D voxel space, and has similar restrictions; in particular, it does not reflect the extent to which the topology of the building models corresponds to the topology of the reference. We based the analysis of the quality of the segmentation on a comparison of roof plane label images, carried out in a way similar to the overlap analysis for the evaluation of object detection, but without topological clarification. We determine several quality metrics to evaluate the quality of the submitted results:

- $Comp_{obj}$, $Corr_{obj}$, Q_{obj}: *completeness*, *correctness* and *quality* determined on a per-roof-plane level. These indices count the number of roof planes that have an overlap of at least 50% with roof planes in the reference. However, no 1:1 relations between roof planes in the reference and roof planes in the reconstruction results are required to count a plane as a true positive. A low completeness indicates that a large proportion of reference planes is not covered by any plane in the extraction results, thus it shows that certain roof parts are not reconstructed at all. A low correctness indicates that many reconstructed roof planes have no correspondence in the reference at all, i.e., they correspond to other objects than roofs. This metric requires an area threshold, and consequently the results will depend on this threshold. See also the discussion in the context of object detection (Rutzinger et al., 2009).
- $Comp_{10}$, $Corr_{10}$ and Q_{10}: *completeness*, *correctness* and *quality* determined on a per-roof-plane level, but only considering roof planes larger than 10 m^2, again to analyse the dependency of these indicators on the object size.
- $N_{1:M}/N_{N:1}/N_{N:M}$: These numbers are related to differences in the topologies of the extracted roof planes and the reference. $N_{1:M}$ is the number of instances where 1:M relations between roof planes in the reference and planes in the reconstruction results occur, thus indicating over-segmentation. Similarly, $N_{N:1}$ counts the number of N:1 relations and is an indicator for under-segmentation. Finally, $N_{N:M}$ is the number of N:M relations between roof planes in the reference and planes in the reconstruction results. It indicates clusters of planes that are both over- and under-segmented. These numbers thus also reflect the quality of the roof plane segmentation.
- *RMSXY*: The geometrical error in *planimetry* was evaluated in a similar way as for object detection. We determined the RMS errors of the planimetric distances of the reference roof plane boundary points to their nearest neighbours on the corresponding extracted roof plane boundaries.
- *RMSZ*: The RMS errors of the height differences are derived by comparing two synthetic DSMs generated from the 3D building models. *RMSZ* is based on the height differences between the reference planes and all corresponding extracted planes. Thus, it also includes a component due to segmentation errors. In a 2.5D setting, disregarding the uncertainties of the ground heights and assuming grid-based processing, negative height errors would correspond to false negative and positive height errors to false positive volumes in the evaluation scheme by Meidow and Schuster (2005). By using the geometrical measure we do not differentiate between these error types, but give an overall measure for the deviations between the two models.

4. Methods

4.1. Task 1: urban object detection

Results of 27 different methods were submitted for this task; an overview can be seen in Table 1. To our knowledge, all methods worked fully automatically in the sense that there was no user interaction in the course of the classification/segmentation/detection steps apart from selecting model parameters or determining training data off-line, and that the output of the method was not subject to any manual intervention such as post-editing. We categorise the methods according to three criteria (note that we write the number of participants that fit into a specific category in brackets behind each definition):

(1) Data used by the participants (column *T* in Table 1): Here we distinguish five groups of methods:
   - *P*: methods solely using the original ALS point cloud as input, sometimes using a Triangulated Irregular Network (TIN) to represent the DSM (8 methods).
   - *D*: methods using an ALS-based DSM grid (3 methods).
   - *DO*: methods combining a grid-based DSM from ALS with an orthophoto (10 methods).
   - *I*: methods solely based on images (4 methods). Note that all of them also use a DSM generated from these images. The criterion is that no ALS data were used.
   - *PI*: methods exploiting the original ALS points and the original images (i.e., no orthophotos as in *DO*) (2 methods).

Table 1
Overview of the detection methods. *T*: Data type used; *ID*: Identifier of the method used in this paper. *Researcher*|*Affiliation*: name and affiliation of the person submitting the results. *Strat:* Detection strategy. *Pr*: Primitives classified in detection. *Res*: type of results submitted by the respective method; *B*: buildings; *T*: trees. *Reference*: a reference where the method is described. Please refer to the main text for an explanation of the abbreviations in columns *T*, *Strat*, *Pr* and *Res*.

T	ID	*Researcher*	*Affiliation*	*Strat*	*Pr*	*Res*	*Reference*
P	UMTA	L. Feng	Univ. Mining Tech., China	*MOD*	*P*	*B*	(ISPRS, 2013)
	UMTP	L. Feng	Univ. Mining Tech., China	*MOD*	*P*	*B*	(ISPRS, 2013)
	MON	M. Awrangjeb	Monash Univ., Australia	*MOD*	*PS*	*B*	(ISPRS, 2013)
	VSK	P. Dorninger	TU Vienna, Austria	*MOD*	*S*	*B*	(Dorninger and Pfeifer, 2008)
	WHUY1	B. Yang	Wuhan Univ., China	*SAM*	*P*	*B*	(Yang et al., 2013)
	WHUY2	B. Yang	Wuhan Univ., China	*SAM*	*P*	*B*	(ISPRS, 2013)
	HANC1	J. Niemeyer	Univ. of Hannover, Germany	*SUP*	*P*	*B* + *T*	(Niemeyer et al., 2011)
	HANC2	J. Niemeyer	Univ. of Hannover, Germany	*SUP*	*P*	*B* + *T*	(Niemeyer et al., 2013)
D	MAR1	D. Mongus	Univ. of Maribor, Slovenia	*MOD*	*S*	*B*	(Mongus et al., 2013)
	MAR2	D. Mongus	Univ. of Maribor, Slovenia	*MOD*	*S*	*B*	(Mongus et al., 2014)
	TON	C. Liu	Tonji Univ., China	*MOD*	*S*	*B*	(Liu et al., 2012)
I	DLR	T. Bucher	German Aerospace, Germany	*MOD*	*S*	*B* + *T*	(ISPRS, 2013)
	FIE	D. Bulatov	Fraunhofer Inst., Germany	*MOD*	*P*	*B*	(Bulatov et al., 2014)
	HAND	F. Rottensteiner	Univ. of Hannover, Germany	*MOD*	*P*	*B*	(Rottensteiner et al., 2007)
	RMA	C. Beumier	Royal Military Ac., Belgium	*MOD*	*P*	*B*	(ISPRS, 2013)
DO	MEL	M. Awrangjeb	Univ. of Melbourne, Australia	*MOD*	*PS*	*B*	(Awrangjeb et al., 2012)
	CAL1	A. Moussa	Univ. of Calgary, Canada	*MOD*	*S*	*B* + *T*	(Moussa and El-Sheimy, 2012)
	CAL2	A. Moussa	Univ. of Calgary, Canada	*MOD*	*S*	*B* + *T*	(ISPRS, 2013)
	LJU1	D. Grigillo	Univ. of Ljubljana, Slovenia	*MOD*	*S*	*B* + *T*	(Grigillo and Kanjir, 2012)
	LJU2	D. Grigillo	Univ. of Ljubljana, Slovenia	*MOD*	*S*	*B* + *T*	(ISPRS, 2013)
	TEH	F. Mahmoudi	Univ. of Tehran, Iran	*MOD*	*S*	*B* + *T*	(ISPRS, 2013)
	KNTU	A. Zarea	K.N. Toosi Univ., Iran	*SUP*	*P*	*B* + *T*	(ISPRS, 2013)
	TUM	W. Yao	TU Munich, Germany	*SUP*	*P*	*B* + *T*	(Wei et al., 2012)
	WHUZ	Q. Zhan	Wuhan Univ., China	*SUP*	*P*	*B* + *T*	(Zhan et al., 2012)
	ZJU	D. Chai	Zejiang Univ., China	*SUP*	*P*	*B*	(ISPRS, 2013)
PI	ITCM	M. Gerke	ITC, The Netherlands	*MOD*	*S*	*B* + *T*	(Gerke and Xiao, 2014)
	ITCR	M. Gerke	ITC, The Netherlands	*SUP*	*S*	*B* + *T*	(Gerke and Xiao, 2014)

(2) Main processing strategy (column *Strat* in Table 1): Here we distinguish three groups of methods:
- *MOD*: approaches are predominately model-based, by which we mean that in some way they are based on an explicit model of the appearance of the objects in the data and that they do not use training (18 methods).
- *SUP*: approaches applying a supervised classification methodology based on training data and some (probabilistic or non-probabilistic) classifier trained on them (7 methods).
- *SAM*: methods based on statistical sampling, using heuristic models for the energy functions (2 methods).

(3) Primitives that are primarily classified (column *Pr* in Table 1):
- *P*: methods classifying image or DSM pixels or ALS points (13 methods).
- *S*: methods classifying image or point cloud segments (12 methods).
- *PS*: methods in which pixel- and point-based processing is equally important (2 methods).

As far as the *results* (column *Res* in Table 1) are concerned, fourteen methods only delivered buildings and thirteen delivered both buildings and trees. Most participants only delivered results for Vaihingen (and some only for one or two of the test areas). For Toronto we only received seven submissions, all of them only delivering building outlines. For lack of space, no descriptions of the methods can be provided here. The reader is referred to the references given in Table 1. Note that reference (ISPRS, 2013), the web site of the test, might only contain the short descriptions supplied by the test participants.

4.2. Task 2: 3D building reconstruction

For this task, results obtained by 14 different methods were submitted (cf. Table 2). We categorised the approaches according to the input data in the same way as the detection methods (cf. Section 4.1). In this case, the vast majority (10) relies on the original ALS points, usually in a TIN (*P*). One method is based on a raster DSM from ALS (*D*) and another one combines the images with the original ALS points (*PI*). Two methods solely rely on images (*I*) in the sense described in Section 4.1. In addition, we categorised the methods according to two criteria:

(1) Degree of automation (column *Au* in Table 2):
- *F*: methods working fully automatically according to the criteria defined in Section 4.1 (12 methods).
- *S*: semi-automatic methods, i.e. methods relying on some human intervention beyond the selection of processing parameters or the generation of training data (2 methods).

(2) Building model applied in reconstruction (column *M* in Table 2):
- *P*: building reconstruction based on primitives (1 method).
- *G*: generic (usually polyhedral) building models (8 methods).
- *A*: The submissions from ITC pursue an adaptive strategy, using pre-defined models but allowing variable influence of these models in the reconstruction process (5 methods).

Again, the reader is referred to the publications cited in the table for more information about the individual methods.

Table 2
Overview of the reconstruction methods. *T*: Data type used; *ID*: Identifier of the method used in this paper. *Researcher*/*Affiliation*: name and affiliation of the person submitting the results. *Au:* level of automation in reconstruction (***F***ully/***S***emi-automatic). *M:* Type of model used for reconstruction: ***G***eneric, ***P***rimitive, ***A***daptive. *Reference*: a reference where the method is described. Please refer to the main text for an explanation of the abbreviations in column *T*.

T	ID	*Researcher*	*Affiliation*	*Au*	*M*	*Reference*
P	MON	M. Awrangjeb	Monash Univ., Australia	*F*	*G*	(ISPRS, 2013)
	VSK	P. Dorninger	TU Vienna, Austria	*F*	*G*	(Dorninger and Pfeifer, 2008)
	ITCE1	S. Oude Elberink	ITC, The Netherlands	*F*	*A*	(Oude Elberink and Vosselman, 2009, 2011)
	ITCE2	S. Oude Elberink	ITC, The Netherlands	*F*	*A*	(Oude Elberink and Vosselman, 2009, 2011)
	ITCX1	B. Xiong	ITC, The Netherlands	*S*	*A*	(Xiong et al., 2014)
	ITCX2	B. Xiong	ITC, The Netherlands	*F*	*A*	(Xiong et al., 2014)
	ITCX3	B. Xiong	ITC, The Netherlands	*F*	*A*	(Xiong et al., 2014)
	CAS	Y. Xiao	Chinese Acad. Sc.	*F*	*G*	(ISPRS, 2013)
	TUD	S. Perera	TU Dresden, Germany	*F*	*G*	(Perera et al., 2014)
	YOR	G. Sohn	York University, Canada	*F*	*G*	(Sohn et al., 2008, 2012)
D	KNTU	A. Zarea	K.N. Toosi Univ., Iran	*F*	*G*	(ISPRS, 2013)
I	FIE	D. Bulatov	Fraunhofer Inst., Germany	*F*	*G*	(Bulatov et al., 2014)
	CKU	J.-Y. Rau	N. Cheng-Kung U., Taiwan	*S*	*G*	(Rau and Lin, 2011)
PI	BNU	W. Zhang	Beijing N. Univ., China	*F*	*P*	(Zhang et al., 2011)

5. Results and discussion

Here we provide a detailed analysis of the results submitted to the benchmark. Note that tables with quality indices for the individual test areas 1–5 are omitted here for lack of space; they can be found on the web site of the test (ISPRS, 2013). Instead we refer to average values for the Vaihingen and the Toronto data.

5.1. Task 1: urban object detection

5.1.1. Building detection

5.1.1.1. Vaihingen: results. Table 3 shows the average quality metrics achieved for the Vaihingen test site. These numbers measure the average building detection performance for three different types of urban development. Note that TEH only delivered results for area 1 and FIE only for area 3. For LJU2 we did not receive results for area 2. All the other methods submitted results for all three areas.

Looking at Table 3, one can see that 13 methods achieve an area-based completeness $Comp_{ar}$ larger than 90%, whereas the correctness $Corr_{ar}$ is larger than 90% for 20 techniques. There are only five methods either having an area-based completeness or an area-based correctness below 85%. The best trade-off between false positives and false negatives is achieved by DLR, ZJU, LJU1 and LJU2, who have similar average quality values $Q_{ar} > 89\%$. For the amount of work that can be saved by automatic building detection, the object-based quality indices may be more relevant. The best method in terms of average object-based completeness is LJU2, but it still misses on average about 12% of the buildings in the three areas. HANC2 is the only other method achieving $Comp_{obj} > 85\%$. Most of the methods deliver very few false positive buildings: six methods do not deliver a single false positive building in any of the three scenes, and $Corr_{obj}$ is larger than 90% for 19 methods. There are only four methods achieving an average quality Q_{obj} larger than 80%, LJU2 being the only one achieving $Q_{obj} > 85\%$. However, there are also methods with $Q_{obj} < 50\%$. In general, the object-based completeness is smaller than the area-based one, which is caused by small buildings missed by all methods. If only buildings larger than 50 m^2 are considered (i.e., 76% of the buildings in Vaihingen), we see that eight methods deliver all such buildings without a single false positive ($Comp_{50} = Corr_{50} = 100\%$). Only three methods miss on average more than 5% of the buildings larger than 50 m^2 ($Comp_{50} < 95\%$), and only two produce more than 5% false positives ($Corr_{50} < 95\%$). The average RMS errors are all larger than 0.60 m, which corresponds to about 1.5 times the average point spacing of the ALS data or to 8 times the GSD of the images. Whereas this would be what one might expect for ALS-based methods, the geometrical potential of the images is certainly not yet fully exploited by any of the image-based methods.

Area 1 provides the most difficult setting for most methods; the area-based quality measures and the geometrical accuracy in general are a bit lower than for areas 2 and 3 (ISPRS, 2013). In area 1 we can identify two common reasons for detection errors (cf. Fig. 1 for some examples). Firstly, none of the methods can detect small buildings. Secondly, all methods seem to have problems with complex roof structures comprising roof parts at different height levels (Fig. 1d). This is particularly the case if some of these roof parts (usually lower appendices to large buildings) are relatively small and/or affected by shadow (Fig. 1b) or in the presence of roof decks. The situation is aggravated by objects on the roof decks (Fig. 1c) or if there are different terrain heights at opposite sides of the roof (Fig. 1a). Some of the methods over-estimate the building size in such scenarios, but most of them will miss some of the building parts. In any case, large height differences both within buildings and at building boundaries seem to make a correct separation of buildings from the terrain difficult. A specific problem of HANC1 and HANC2, solely based on ALS points, is the generation of a relatively high false positive rate due to a confusion of trees having smooth canopies with buildings (Fig. 1e). MAR1 and MAR2, both based on morphological profiles, seem to have problems with inner courtyards (Fig. 1f).

The most favourable conditions are encountered in area 2, although the object-based quality measures are affected by the low number of buildings in that area. This is the area where the highest percentage of methods can detect all buildings larger than 50 m^2 without producing a false positive of that size ($Comp_{50} = Corr_{50} = 100\%$) (ISPRS, 2013). It would seem that in area 2, the clearer structure of the building outlines compared to area 1 has a positive impact on the geometrical quality of the results, indicated by RMS values that are better than the average. We can identify three major error sources in area 2, two of them affecting all methods. As in area 1, small buildings cannot be detected reliably by any method. Secondly, there is one flat roof part covered by vegetation that is not detected by any of the methods (cf. Fig. 2a). It is also very low, being the roof of a basement whose floor is below the terrain level. Thirdly, some of the methods produce a rather large rate of false positive buildings that actually correspond to high trees. The reasons for this may be different depending on the specific situation or on the method: sometimes the error seems to be related to shadows and matching errors next to buildings (e.g. HAND; cf. Fig. 2b), in other cases it is caused by the smooth

Table 3
Evaluation of the building detection results in Vaihingen: average of areas 1–3. The quality metrics are explained in Section 3.1. Data types (*T*) and identifiers (*ID*) are identical to those in Table 1. The best values per column are printed in bold font.

T	*ID*	Area-based (%)			Object-based (%)			Object-based (50 m^2) (%)			*RMS* (m)
		$Comp_{ar}$	$Corr_{ar}$	Q_{ar}	$Comp_{obj}$	$Corr_{obj}$	Q_{obj}	$Comp_{50}$	$Corr_{50}$	Q_{50}	
P	UMTA	92.3	87.5	81.5	80.0	98.6	79.1	99.1	**100.0**	99.1	0.87
	UMTP	92.4	86.0	80.3	80.9	95.8	78.1	98.8	97.2	96.0	0.97
	MON	92.7	88.7	82.8	82.7	93.1	77.7	99.1	**100.0**	99.1	0.93
	VSK	85.8	**98.4**	84.6	79.7	**100.0**	79.7	97.9	**100.0**	97.9	0.87
	WHUY1	87.3	91.6	80.8	77.6	98.1	76.5	97.4	97.9	95.4	0.83
	WHUY2	89.7	90.9	82.3	83.0	97.5	81.3	99.1	98.0	97.2	0.90
	HANC1	91.5	92.5	85.2	81.5	72.7	62.4	**100.0**	95.8	95.8	0.67
	HANC2	90.2	93.2	84.6	85.1	69.6	61.9	**100.0**	**100.0**	**100.0**	0.83
D	MAR1	87.0	97.1	84.8	78.2	96.2	75.7	99.1	**100.0**	99.1	0.83
	MAR2	89.7	95.2	85.8	80.6	93.7	76.5	99.1	98.9	98.0	0.83
	TON	77.7	97.7	76.3	67.5	98.9	66.9	92.7	98.8	91.6	0.90
I	DLR	93.3	96.0	**89.8**	80.3	99.0	79.6	**100.0**	**100.0**	**100.0**	0.73
	FIE	89.0	86.9	78.5	78.6	**100.0**	78.6	**100.0**	**100.0**	**100.0**	1.20
	HAND	93.6	90.3	85.0	80.3	88.8	73.0	97.4	97.2	94.6	0.83
	RMA	92.8	90.2	84.2	82.7	81.0	68.1	**100.0**	**100.0**	**100.0**	0.90
DO	MEL	88.0	79.2	71.4	75.9	76.1	59.7	97.4	81.3	78.8	1.10
	CAL1	89.8	95.1	85.8	76.2	**100.0**	76.2	96.5	**100.0**	96.5	0.73
	CAL2	89.2	97.2	87.0	78.2	**100.0**	78.2	**100.0**	**100.0**	**100.0**	0.77
	LJU1	94.2	94.6	89.4	83.0	**100.0**	83.0	**100.0**	**100.0**	**100.0**	0.73
	LJU2	**94.6**	94.4	89.5	**87.9**	**100.0**	**87.9**	**100.0**	**100.0**	**100.0**	0.75
	TEH	76.7	93.8	73.0	75.7	90.3	70.0	85.7	**100.0**	85.7	1.00
	KNTU	87.7	93.5	82.6	80.9	93.4	76.5	**100.0**	**100.0**	**100.0**	0.93
	TUM	89.7	92.9	83.9	80.9	99.0	80.2	99.1	**100.0**	99.1	1.03
	WHUZ	80.3	89.5	73.2	66.6	55.0	42.0	83.6	95.7	80.7	1.10
	ZJU	92.8	96.4	89.7	76.4	97.0	74.8	99.1	**100.0**	99.1	**0.63**
PI	ITCM	92.7	80.9	75.9	84.8	51.2	47.1	99.1	88.9	88.0	1.13
	ITCR	91.4	90.6	83.5	80.0	70.6	60.0	98.2	**100.0**	98.2	0.93
Average areas 1–3 (107 buildings)											

appearance of tree canopies in the ALS point clouds (e.g. HANC1, HANC2, similar as in area 1).

In area 3, the area-based quality indices and the RMS errors are similar to area 2, but it has the lowest percentage of methods achieving $Comp_{50} = Corr_{50} = 100\%$ (ISPRS, 2013). This seems to be related to the fact that in this suburban setting, the separation between vegetation (which is more abundant than in area 1) and buildings (which are not as clearly structured as in area 2) becomes difficult. As in areas 1 and 2, small buildings are missed by all methods, but here the situation is aggravated by vegetation next to these buildings (Fig. 3a). Large trees next to buildings create problems for some methods based on the segmentation of the DSM (Fig. 3b). The fact that the data were collected under leave-on conditions also leads to an increased false positive rate for some of the ALS-based techniques. Techniques based solely on images do also use DSMs, in this case derived by image matching, and the quality of the matching results directly affects building detection. Fig. 3c shows a typical example where a DSM error leads to a missed building. Though this specific error only occurs with HAND (which uses its own technique for generating a DSM), the other purely image-based methods have similar problems in areas where their matcher delivers wrong results. Fig. 3d shows a building with a very low ALS point density that is missed by most of the ALS-based methods and also by some methods that combine ALS and image data. The situation can be overcome by methods based on a contextual classification such as HANC2.

The individual methods have problems with different situations, all of them having their own strengths and weaknesses. Nearly all situations that are critical for some techniques are correctly resolved by at least one other method, with two exceptions: Firstly, all methods fail to deliver very small buildings, so this seems to be a common weakness of current building detection techniques. Secondly, all methods fail with the situation in Fig. 2a, that is with the roof of a basement that is also covered by vegetation. This is a rather tricky situation, because the reflecting material actually is "low vegetation", but in the context of building detection one would expect this region to be classified as a building part. Current methods do not seem to be able to cope with such ambiguous situations. Nevertheless one can say that most state-of-the-art methods are suitable for detecting the largest buildings on a block of land very reliably. However, the geometrical quality of the outlines is somewhat questionable. The RMS errors are relatively large in general, and there are also outliers and very irregular shapes of some building outlines that would require manual post-editing even for topographic databases corresponding to medium mapping scales. Applications such as updating of the cadastre, requiring a very high accuracy and a very high level of detail and, thus, also the detection of small buildings, do not seem to be feasible in a fully automatic way yet.

5.1.1.2. Vaihingen: comparison of the potential of different input data sets and processing strategies. An interesting question arising from this test is whether there are differences in the potential of different input data sets and processing strategies for urban object detection. Table 4 shows the average quality metrics achieved for all methods using the same input data sets, along with their standard deviations. An important point is that the quality of the results shows more variation within each data set group than between the groups. For instance, the average area-based quality varies between 79.7% for methods using the ALS points and the original images (*PI*) and 84.4% for methods combining a DSM grid from ALS and an orthophoto (*DO*). The difference between these values is 4.7%, but the variation of the quality of the results within the respective data set groups, expressed by the standard deviations in Table 4, is 5.3% and 4.6%, respectively. Therefore, the difference between groups relying of different input data sets is hardly

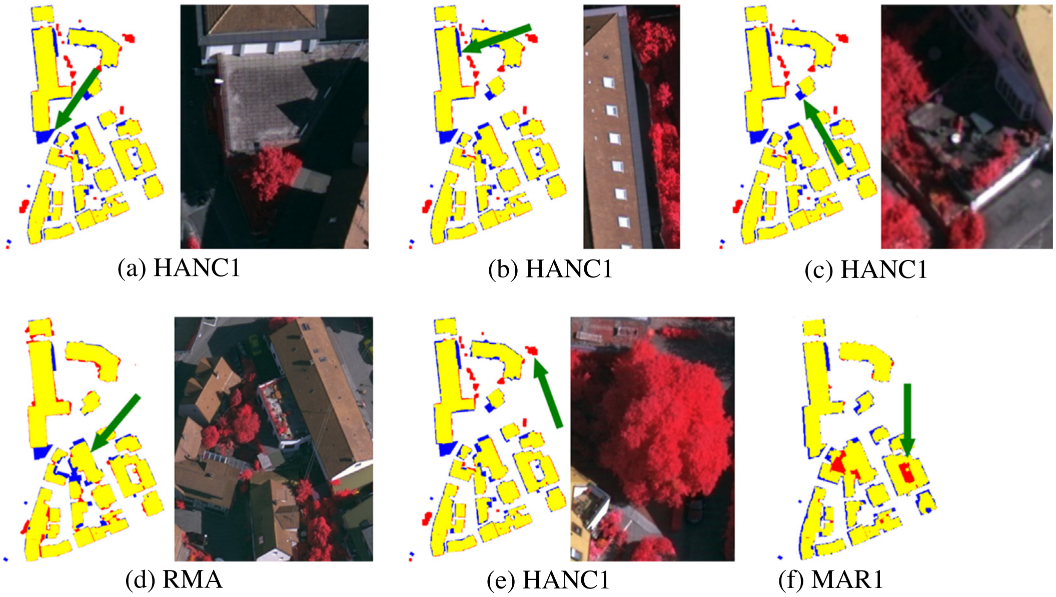


Fig. 1. Typical situations causing errors in area 1. (a) Horizontal roof-deck with different terrain height levels to the east and to the west. (b) Small roof appendix in the shadow of trees. (c) Roof deck with objects on top of it. (d) Complex shape, including many small roof planes and a roof deck. These errors also occur with other methods than those shown in the figure. Specific problems of some methods: (e) Dense tree canopies (HANC1, HANC2). (f): Inner courtyards (MAR1, MAR2).

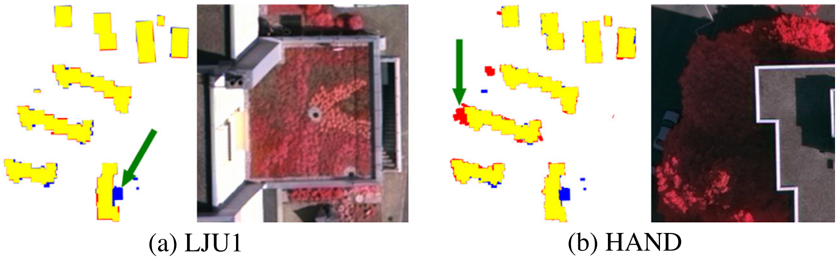
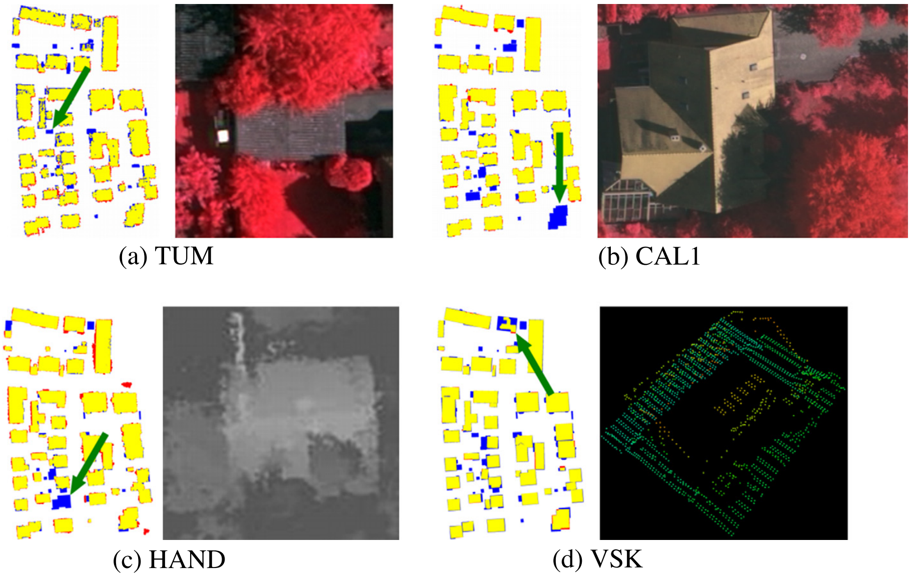


Fig. 2. Typical situations causing errors in area 2. (a): Roof of a basement covered by vegetation (not detected by any method). (b) A tree in the shadow causing problems for the image-based method HAND.

significant. The same applies to the other quality measures, including the geometrical errors. Most of the six methods achieving the highest average area-based quality Q_{ar} use a DSM grid from ALS and an orthophoto (*DO*), but this list also includes methods based on images (*I*) and a DSM grid from ALS (*D*). The six techniques achieving the lowest average area-based quality Q_{ar} also include methods based on combinations of ALS and image data (*DO* and *PI*), on a DSM from ALS (*D*) and on images (*I*). It would seem that this comparison remains inconclusive: methods that do include images may have a slight advantage over methods not using them in terms of the area that is classified correctly, but the quality of the results seems to be much more influenced by the specific technique that is used by the participants. More specifically, the problems still encountered in building detection today are such that the advantage of images in terms of GSD and radiometric content does not seem to give a method a crucial advantage that could not be achieved without that data source. In general, the geometrical accuracy is on the level that can be expected for methods based on ALS data for the best methods. It does come as a bit of a surprise that using images of 8 cm GSD does not lead to an improvement of the geometry of the resulting models. This may be due to the fact that in all cases except ITCM and ITCR an orthophoto was the basis of the image information. We do not know how the orthophoto was generated in most cases; if generated on the basis of a DTM, it would contain systematic displacements of the roof outlines, whereas if based on a DSM, it would be blurred due to geometrical errors of the DSM at building outlines. Using the original images might be a strategy to improve the geometrical quality of building outlines, but it would require more sophisticated processing techniques as they are applied in 3D reconstruction. Generally, the results of our test suggest that none of the methods fully exploits the information contained within the optical images of 8 cm GSD.

We also compared the average quality metrics achieved for Vaihingen for different processing strategies (Table 5). We distinguish methods that are primarily supervised, model-based, or based on statistical sampling (cf. Section 4.1). Again, the variations within each group are much larger than the variations between the groups. The best five methods in terms of area-based quality (Q_{ar}) comprise four model-based and one supervised technique, whereas among the methods achieving the five lowest Q_{ar} values there are two supervised and three model-based ones. This comparison is also inconclusive based on the information that is accessible to us. The three general strategies seem to have a similar quality potential in building detection. We do not know the amount of time required for parameter tuning in model-based methods, or the amount of work required for generating the training data in supervised methods, considerations that may still lead to preferring one strategy over the other. The second aspect of the processing strategy is whether the method is based on a classification of individual pixels/points or on an initial segmentation (e.g., of the DSM). Table 6 gives average quality metrics for Vaihingen for methods primarily classifying pixels or points (*P*), segments



(a) TUM (b) CAL1

(c) HAND (d) VSK

Fig. 3. Typical situations causing errors in area 3. These errors also occur with other methods than those shown in the figure. (a): Small building, interacting with vegetation. (b) High vegetation next to a building (missed by segment-based methods CAL1, TON). (c) A matching error leading to a missed building in HAND. (d) A roof with very few returns that is missed by several ALS-based techniques.

(*S*), or combining both aspects (*P* + *S*). Again, the differences between the groups are smaller than the variations within each group. Looking at the five methods having the highest Q_{ar} values, segment-based methods seem to dominate, but there is also one pixel-based technique. The methods achieving the five lowest Q_{ar} values include techniques from each of the three groups.

5.1.1.3. Toronto: results. The average quality measures for the Toronto data set (areas 4 and 5) are shown in Table 7. Results for the individual areas can be found in (ISPRS, 2013). Note that FIE only delivered results for area 4. All the other methods listed in Table 7 submitted results for both areas in Toronto.

In Toronto, the area-based metrics are in a similar range as in Vaihingen. Four methods achieve both a completeness and a correctness larger than 90% and, consequently, also Q_{ar} values larger than 85%. In general, there are fewer small buildings in both areas: in area 4, only 2% and in area 5 only 8% of the buildings are smaller than 50 m^2. Consequently, the difference between object-based and area-based completeness is not as high as in Vaihingen and the object-based quality measures are in general somewhat larger. Despite being in the CBD of Toronto, the number of buildings smaller than 50 m^2 is larger in area 5 than in area 4. As many of these small buildings are missed by all methods, slightly smaller object-based completeness values are achieved in area 5 compared to area 4 (ISPRS, 2013). Three methods for which results were submitted for both Toronto areas achieved an average object-based quality larger than 80% (FIE achieved 96.6%, but only submitted results for the simpler scenario in area 4). ITCR and ITCM deliver many false positives, but most of them are smaller than 50 m^2. As a consequence, these two methods are the only ones who achieve much better correctness for objects larger than that threshold ($Corr_{50}$). For all other methods, the differences between quality measures for all objects and those achieved considering only buildings larger than 50 m^2 are relatively small, the largest proportion of the effect occurring with area 5 for the reason explained above. The RMS errors are slightly larger than in Vaihingen (0.8–2.1 m). The mismatch between ALS and images, caused by the fact that the datasets were acquired in different vegetation periods, explains the large errors from ITCR and ITCM because those methods rely on good colour information for the separation of sealed and vegetated areas. The RMS errors are in a similar range in both areas, except for MAR2, which delivers results that are better by a factor 2 in area 5 than in area 4 (ISPRS, 2013).

Table 4
Average quality metrics and their standard deviations as a function of the original input data for building detection (areas 1–3). *P*: ALS points or TIN; *D*: DSM grid from ALS; *I*: images only; *DO*: DSM grid from ALS and orthophoto; *PI*: original ALS points and images. The numbers in parenthesis are the number of submissions per group.

	P (8)	*D* (3)	*I* (4)	*DO* (10)	*PI* (2)
$Comp_{ar}$ (%)	90.3 ± 2.6	84.8 ± 6.3	92.2 ± 2.1	88.3 ± 5.8	92.1 ± 0.9
$Corr_{ar}$ (%)	91.1 ± 3.9	96.7 ± 1.3	90.9 ± 3.8	92.7 ± 5.2	85.7 ± 6.9
Q_{ar} (%)	82.8 ± 1.9	82.3 ± 5.2	84.4 ± 4.6	82.6 ± 7.3	79.7 ± 5.3
$Comp_{obj}$ (%)	81.3 ± 2.3	75.5 ± 7.0	80.5 ± 1.7	78.2 ± 5.6	82.4 ± 3.4
$Corr_{obj}$ (%)	90.7 ± 12.2	96.3 ± 2.6	92.2 ± 9.0	91.1 ± 14.7	60.9 ± 13.8
Q_{obj} (%)	74.6 ± 7.8	73.0 ± 5.3	74.8 ± 5.3	72.8 ± 13.2	53.5 ± 9.1
$Comp_{50}$ (%)	98.9 ± 0.9	97.0 ± 3.7	99.3 ± 1.3	96.1 ± 6.2	98.7 ± 0.6
$Corr_{50}$ (%)	98.6 ± 1.6	99.2 ± 0.7	99.3 ± 1.4	97.7 ± 5.9	94.5 ± 7.8
Q_{50} (%)	97.6 ± 1.7	96.2 ± 4.1	98.7 ± 2.7	94.0 ± 8.7	93.1 ± 7.2
RMS (m)	0.86 ± 0.09	0.86 ± 0.04	0.92 ± 0.20	0.88 ± 0.17	1.03 ± 0.14

Despite the fact that the percentage of small objects is lower in Toronto than in Vaihingen, the problems in detecting small objects exist as well. Unlike in Vaihingen, no method is capable of detecting all buildings larger than 50 m^2 or delivering no false positives larger than 50 m^2. Partly this may be attributed to the fact that the data set in Toronto is larger, but there are also some specific problems. The first major problem is occlusion (in case images are used; cf. Fig. 4a), which is more prominent here due to the lack of a multi-image configuration and a much higher parallax range. There are also problems with buildings having very different height levels, but in addition there is a much larger variation in the average building height, which may be the reason why the relatively low buildings in Fig. 4b and c are missed by some supervised methods. In Fig. 4c, the situation is aggravated by the shadow, which gives this low building a rather different radiometric appearance than the other ones. It is worth noting that despite the challenges for image matching in this scenario, one method (FIE) is based entirely on images and produces rather good results in area 4. Again, the geometrical potential of image data is not exploited by any of the methods, the average RMS errors being

Table 5
Average quality metrics and their standard deviations as a function of the processing strategies for building detection (areas 1–3). The numbers in parenthesis are the number of submissions per group.

	Supervised (*SUP*) (7)	*Model-based* (*MOD*) (18)	*Sampling* (*SAM*) (2)
$Comp_{ar}$ (%)	89.3 ± 4.4	89.5 ± 5.1	88.5 ± 1.7
$Corr_{ar}$ (%)	91.3 ± 5.0	92.2 ± 5.1	91.2 ± 0.5
Q_{ar} (%)	82.2 ± 5.7	83.0 ± 5.3	81.6 ± 1.0
$Comp_{obj}$ (%)	79.5 ± 6.4	79.4 ± 4.2	80.3 ± 3.8
$Corr_{obj}$ (%)	76.9 ± 19.9	93.5 ± 9.0	97.8 ± 0.4
Q_{obj} (%)	63.5 ± 14.8	74.9 ± 7.4	78.9 ± 3.4
$Comp_{50}$ (%)	97.3 ± 6.0	97.8 ± 3.5	98.3 ± 1.2
$Corr_{50}$ (%)	97.2 ± 4.2	98.5 ± 4.4	98.0 ± 0.1
Q_{50} (%)	94.7 ± 7.5	96.4 ± 5.7	96.3 ± 1.3
RMS (m)	0.90 ± 0.20	0.88 ± 0.13	0.87 ± 0.05

5–14 times larger than the GSD of 15 cm; the best ALS-based techniques are in the order of three times the average point spacing, thus coming closer to the inherent potential of these data.

5.1.1.4. Building detection: discussion. In their paper on the EuroSDR test for road extraction, Mayer et al. (2006) claimed that a completeness of 70% and a correctness of 85% are required for object detection techniques to become of practical importance. Our evaluation shows that these quality measures can be achieved by most of the building detection techniques. It would seem that automated building detection for topographic mapping is relatively mature and can deliver very reliable results for the main buildings on a plot of land, at least if only the existence of a building is to be shown. The situation is much less favourable if small buildings are of interest or if a very precise delineation of the buildings is required, as it would be the case for large scale maps such as the cadastre. A comparison of the use of different input data or processing techniques remained inconclusive. However, many of the compared methods did require an NDVI, so they could not be applied to the Toronto data. We have to note that some methods delivered rather good results for Toronto given the reduced resolution and the fact that the images did not contain an infrared band, perhaps because height is rather discriminative there.

The question remains where to spend additional research to further advance the state-of-the-art of building detection beyond what can already be achieved by the best methods compared in this paper. As all methods had problems with small buildings independently from the data set, this would be one obvious direction. Here, methods based on ALS in the resolution given in this benchmark may reach their limits, but image-based methods should be capable to deliver this information given a GSD of 8 or 15 cm. It would seem that one of the bottlenecks is the matching required for obtaining a DSM: in the presence of wide baselines and object planes that are not nearly parallel to the image planes, but also in shadow areas, state-of-the-art dense matching still encounters problems. Thus, work to improve dense matchers for such scenarios might also improve the prospects of building detection. An analysis involving multiple images might help as well, in particular if they include views from oblique angles: if information about the walls were included in the process (which is not done by any of the compared methods), there might be more direct cues for detecting small building structures. This would involve a high-level processing stage in which the results from multiple views are merged. Such a strategy might also help to solve ambiguous situations with vegetation on roofs. In order to improve the geometrical quality of the outlines, it would seem to be necessary to include the original aerial images into the process. The required processes could, for instance, try to detect 3D lines by matching image lines from multiple images, comparing these lines with the coarse outlines delivered by current methods. Finally, different strategies for primary data acquisition may also help to advance the state-of-the-art. Multiple-overlap imagery is required in densely built-up areas, anyway, in order to avoid occlusions. In addition, the potential of oblique airborne imagery for automated object detection still remains to be investigated.

Table 6
Average quality metrics and their standard deviations as a function of the primitives that are classified in building detection (areas 1–3). *P*: ALS points or (DSM or image) pixels; *S*: Segments; *P* + *S*: Pixels and segments. The numbers in parenthesis are the number of submissions per group.

	P (13)	*S* (12)	*P* + *S* (2)
$Comp_{ar}$ (%)	90.0 ± 3.5	91.4 ± 3.0	90.4 ± 3.3
$Corr_{ar}$ (%)	90.9 ± 2.9	91.9 ± 6.2	84.0 ± 6.7
Q_{ar} (%)	82.5 ± 4.0	84.5 ± 5.7	77.1 ± 8.1
$Comp_{obj}$ (%)	79.6 ± 4.5	81.1 ± 3.9	79.3 ± 4.8
$Corr_{obj}$ (%)	88.2 ± 14.3	88.2 ± 19.5	84.6 ± 12.0
Q_{obj} (%)	71.7 ± 11.0	72.5 ± 14.1	68.7 ± 12.7
$Comp_{50}$ (%)	98.0 ± 4.4	99.2 ± 0.9	98.3 ± 1.2
$Corr_{50}$ (%)	98.6 ± 1.7	98.4 ± 4.2	90.7 ± 13.2
Q_{50} (%)	96.7 ± 5.2	97.6 ± 4.3	89.0 ± 14.4
RMS (m)	0.90 ± 0.16	0.89 ± 0.21	1.02 ± 0.12

5.1.2. Tree detection

This section is entirely based on the Vaihingen data because no results for tree detection were submitted for Toronto so far.

5.1.2.1. Results. The average evaluation results for areas 1–3 are shown in Table 8. For TEH we only received results for area 1, whereas all other participants submitted results for all three areas. This analysis will concentrate on the object-based metrics, because they are less affected by generalisation errors in the reference than the area-based ones. The RMS errors of all methods are in the order of 1.3–1.6 m.

Looking at Table 8, one can see that tree detection has a much lower success rate than building detection. Only five methods achieve both object-based completeness and correctness values larger than 50%. The best trade-off and, thus, the best object-based quality Q_{obj} is achieved by KNTU, but it is still below 50%. CAL2 and TUM achieve similar values in the order of 48–50%. The situation is better for trees larger than 50 m^2, though the improvement is smaller than in case of buildings. In this case, TUM clearly outperforms the other methods, detecting 93.2% of the trees larger than 50 m^2 while delivering very few (1.7%) false positives. These numbers clearly show the limitations of current methods for tree detection. If we adopt the criteria for practical relevance by Mayer et al. (2006), we see that eight methods can be relevant for detecting trees larger than 50 m^2. However, trees smaller than that cannot be detected reliably by any of the methods compared. This is even more critical because only about 13% of the trees in the Vaihingen data set have a crown larger than 50 m^2, so that the problems of current methods affect the majority of the trees in such scenes.

In all areas, crown size is an important factor leading to detection errors. As the proportion of small trees is larger than the proportion of small buildings in a scene, this heavily affects the performance of all compared methods. In area 1 an additional reason for failure is the vicinity of complex terrain, causing the failure of several methods to detect the trees indicated by the arrow in Fig. 5a. HANC1 and HANC2 additionally suffer from a confusion of trees and buildings discussed in Section 5.1.1. The results for area 2 are somewhat better than those for areas 1 and 3 because it contains larger trees (ISPRS, 2013). In addition to the error sources described for area 1, rows of bushes separating neighbouring plots of land frequently lead to false positives (Fig. 5b). Area 3 is

Table 7
Evaluation of the building detection results in Toronto: average of areas 4–5. The quality metrics are explained in Section 3.1. Data types (*T*) and identifiers (*ID*) are identical to those in Table 1. The best values per column are printed in bold font.

T	*ID*	Area-based (%)			Object-based (%)			Object-based (50 m^2) (%)			*RMS* (m)
		$Comp_{ar}$	$Corr_{ar}$	Q_{ar}	$Comp_{obj}$	$Corr_{obj}$	Q_{obj}	$Comp_{50}$	$Corr_{50}$	Q_{50}	
P	WHUY2	94.3	91.3	86.5	90.4	95.8	86.8	94.8	95.8	90.9	1.10
D	MAR1	96.1	92.1	88.7	**98.7**	86.8	86.0	98.6	87.6	86.7	**0.80**
	MAR2	94.0	94.3	**88.9**	91.3	91.9	84.8	95.7	96.8	92.7	2.10
I	FIE	**96.6**	90.6	87.8	98.3	**98.2**	**96.6**	**100.0**	**98.2**	**98.2**	1.20
DO	TUM	85.1	80.6	70.6	83.9	90.3	77.0	88.2	92.5	82.3	1.60
PI	ITCR	75.0	**94.5**	71.7	79.6	43.5	38.5	83.8	91.8	77.8	1.05
	ITCM	76.9	87.5	68.9	86.5	21.7	21.0	89.7	70.5	64.8	1.45
Average areas 4–5 (96 buildings)											

a very complex scenario for tree detection, with a large variety of trees of different heights and sizes, and including many small trees. Problems in distinguishing high and low vegetation (e.g. Fig. 5c) are more dominant here than in the other two areas, whereas there are no problems due to complex terrain structures.

5.1.2.2. Comparison of the potential of different input data sets and processing strategies. We compared the average quality numbers for methods based on the same input data (Table 9), and again, the results are inconclusive, the variations between the individual methods of the four groups being in the same order of magnitude as the differences between the average metrics of the groups. We do not exactly know why the problems described above actually do cause the failure of the methods. We suspect that when using images, matching becomes very difficult with trees, in particular if the canopy has a complex structure or if the tree is small. For methods combining ALS and image data, data alignment may become problematic, and for methods solely based on ALS data the point density might just not be high enough to reliably detect small trees. In the latter case, confusion between trees and buildings due to smooth tree canopies sometimes even prevents the detection of larger trees.

Similar comparisons were carried out for the average quality numbers for groups using a similar processing strategy (Table 10) and to compare pixel-based and segment-based approaches (Table 11). However, again the comparison remains inconclusive. Given the evaluation results for Vaihingen, it is impossible to single out a processing strategy to be more promising than the others; the quality of the results seems to depend on the particular method involved. In case of tree detection, the method by TUM, based on boosting, seems to be the one delivering the best results.

5.1.2.3. Tree detection: discussion. Our test indicates that the automatic detection of small trees is a problem that is largely unsolved and still needs to be tackled by future work. On the one hand, improved classifiers taking into account more features than those used in the test could help. The contextual classifiers used in this text only used context on a very local level. Sampling techniques may be a way to improve the situation, because they can propose hypotheses based on a model for the distribution of objects in a scene; such models could include information about the typical alignment of trees in built-up regions, e.g. along property lines (which might again be aligned with roads or buildings). With ALS data, the inclusion of full waveform analysis might also improve the prospects of this task.

5.2. Task 2: 3D building reconstruction

5.2.1. Vaihingen: results

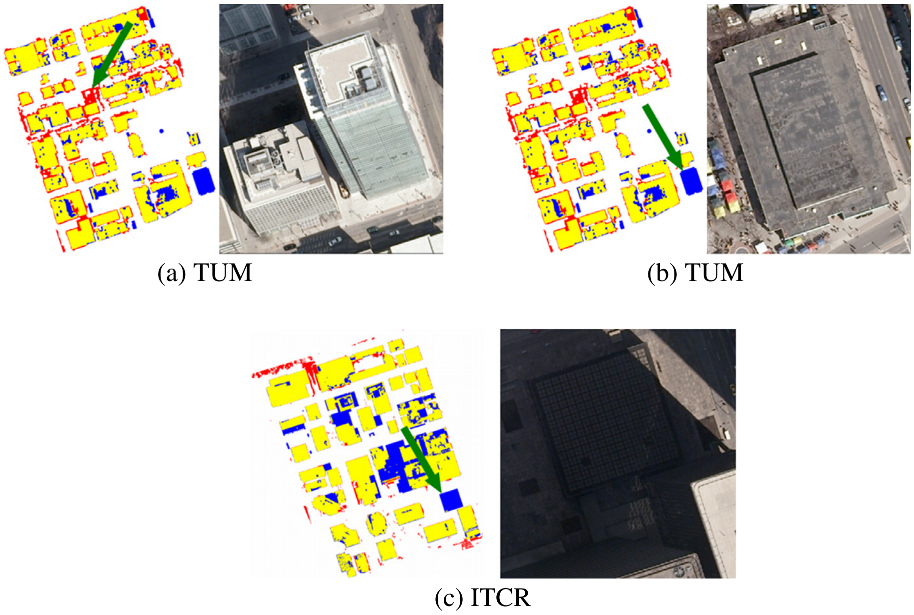
The evaluation of the building reconstruction results for Vaihingen is summarised in Table 12. It gives the average quality indices for the areas 1–3. Again, the individual tables for the three test areas are available on the website of the benchmark (ISPRS, 2013).

Overall, 14 participants submitted 3D building reconstruction results for Vaihingen. Most of them processed all areas. CAS only provided results for areas 2 and 3 and FIE, BNU and KNTU concentrated on area 3.

Whereas the roof-plane based completeness shown in Table 12 shows larger variations between 68.5% (CAS) and 82.8% (ITCX3), the correctness is above 90% for all methods except for FIE and MON. If only large roof planes (>10 m^2) are considered, the completeness rises significantly for most methods, only for ITCX1 it remains below 75%. This is a first indication that small roof planes remain undetected whereas bigger ones are nicely extracted. However, the correctness increases only marginally for most approaches, which means that if wrong planes are detected, they are very likely to be larger than 10 m^2. Concerning the quality of the derived topology we can clearly observe that under-segmentation ($N_{N:1}$) is the dominant error type, that is, planes are merged. However, as we will see below, the topologic quality varies substantially between the areas indicating that the type of roof architecture has a significant impact on the results.

The geometric accuracy in the *XY* plane is worse than the RMSE value in *Z* direction for all methods, but overall in the same range: about 70 cm in planimetry and 30 cm in height. Interestingly, this observation is also valid across methods using different input data. More precisely, methods using solely ALS data achieve a RMSE in *XY* on the same accuracy level as those based on images although the latter samples the ground much more densely. Generally, 70 cm amounts to a bit less than twice the ALS point spacing, indicating a high level of maturity of the respective methods. In comparison, the aerial images have a GSD of about 8 cm and it thus seems that not all information contained in this very densely sampled data is exploited by the corresponding methods yet. In the following, we discuss each test area separately because they comprise different roof types, leading to specific challenges for the reconstruction methods.

Test area 1 is quite challenging for roof reconstruction approaches because various roof structures occur, because buildings are somewhat irregularly distributed, and because there are quite a few very small roof parts. Nonetheless, all ten submissions achieve correctness values of roof planes better than 83% and the majority even shows a correctness higher than 96% (ISPRS, 2013). Note that the roof-based completeness shows much larger variations than correctness, which is a hint that most participants have


(a) TUM

(b) TUM

(c) ITCR

Fig. 4. Typical situations causing errors in areas 4 and 5. These errors also occur with other methods than those shown in the figure: (a) Area 4: occlusion by a skyscraper. (b) Area 4: a building with a non-representative height and two different height levels. (c) Area 5: another rather low building, additionally affected by the shadow of a nearby skyscraper.

tuned their approaches towards achieving a high correctness. The most complete extraction is given by ITCX3 and YOR (89%) whereas ITCE1 misses almost 40% of the roof planes. Under-segmentation is the dominant type of topologic error, occurring in 26–42 cases ($N_{N:1}$) which does not come as a surprise considering the rather complex roof structures in this area. The situation hardly changes regarding the quality metrics that take into account only roof planes larger than 10 m^2. This allows inferring that small roof structures are not a hurdle per se but complex roof architectures are. The best RMS in *XY*-plane (0.8 m) is achieved by TUD and YOR, and has to be seen in relation to the ALS point spacing. Not surprisingly, the height errors (*RMSZ*) are smaller than the planimetric ones for methods based on ALS, and for this area also the image-based method CKU shows better height accuracy.

Compared to area 1, roof reconstruction in area 2 seems less challenging due to several large flat-roofed buildings and less complex roof structures in general. There is a clear difference between the quality metrics for all planes and for roof planes larger than 10 m^2 for all eleven submissions. It would seem that except for ITCE1, most large roof planes can be detected and nearly all of them are correct. Fewer instances of under-segmentation ($N_{N:1}$ values between 3 and 7) occur, possibly because the roofs have less small structures compared to area 1 as already expected. However, for both variants of ITCE, all high-rise buildings, which do have several superstructures, are reconstructed by single planes, demonstrating that this approach has major problems with complex flat roof buildings. The other methods produce quite good reconstructions of the main roof structures; ITCX1 produces a few larger false positive planes at one of the few smaller residential buildings in this area. Over-segmentation is quite dominant at large flat-roof building structures for ITCX2, i.e. without automatic dictionary-based error correction. But once this technique is applied (ITCX3), those errors are mostly corrected. In general, both the planimetric and the height accuracy are slightly worse than in area 1. The main reason for this is that roof boundaries are not computed as plane intersection as in slanted roofs, but estimated from the single planes.

Similar to area 1, area 3 presents some challenges to reconstruction because various buildings with gabled roofs and small superstructures are present in the scene. Consequently, completeness and correctness values are distributed in a similar way as in area 1 across all 14 submissions. Under-segmentation occurs more frequently than in area 1, which may be explained by a large number of small attachments to the houses that are erroneously merged with neighbouring roof planes. On the other hand, and possibly because of the same reason, over-segmentation is hardly evident. YOR achieves the overall best results. Two groups using images only submitted results for this area. Of these groups, CKU shows quite low completeness values, missing some attachments.

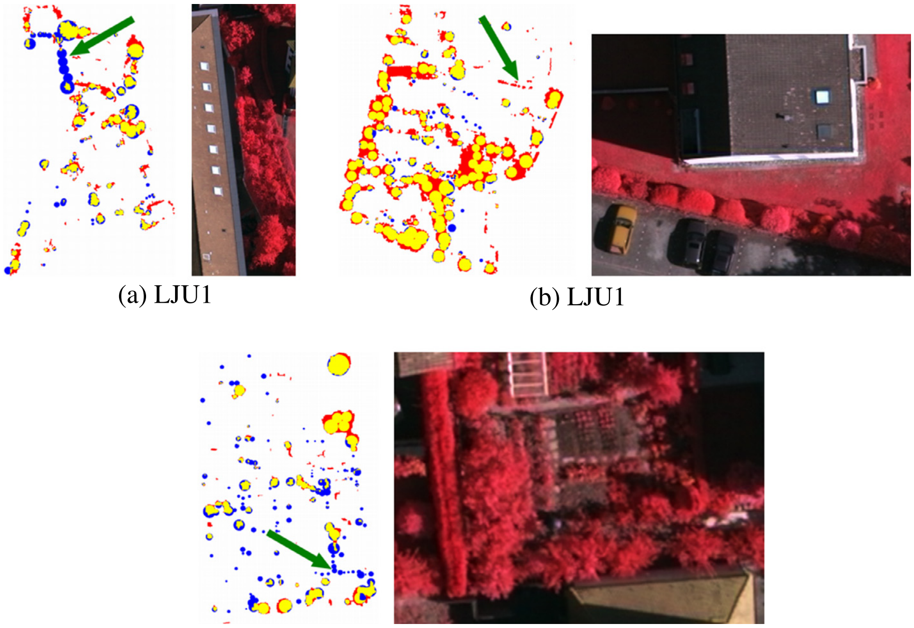
The only method which uses primitive-based reconstruction, BNU, performs quite well, but has many under-segmentation errors, mostly because small superstructures on roofs are not modelled adequately. FIE does a good job in detecting planes, but shows the worst planar RMSE values. Generally, the geometrical errors (*RMS*, *RMSZ*) are in the same range as in area 1.

5.2.1.1. Detailed analysis. The different architectural properties of the three areas allow us to draw several conclusions from the individual results. Areas 1 and 3 are similar in the sense that inclined roofs with small structures such as chimneys are dominant. In both areas, the completeness and correctness values taking into account all planes are quite similar to those considering only the planes larger than 10 m^2. This observation, in conjunction with the relatively large number of under-segmentation errors, indicates that the presence of small building structures may prevent the correct detection of the dominant planes. For example, Fig. 6a shows some missing smaller planes, e.g. the sides of the dormer with the characteristic triangular roof plane, but also missing parts of the two main planes representing the generalised shape of the building, which could be considered to be a saddleback roof. This is a typical example indicating that the problems of current methods with complex roof structures do not just result in more generalised models without details, but may lead to models of more irregular shapes than the reference.

Fig. 6b shows the under-segmentation problem of ITCE1 in area 2. Of course, if the entire roof structure is reconstructed by a single horizontal plane, the height errors will increase. Fig. 6c shows an example for over-segmentation of an ALS point cloud: obviously the points near the ridge of the buildings (where the surface nor-

Table 8
Evaluation of the tree detection results in Vaihingen: average of areas 1–3. The column headings are explained in Section 3.1. Data types (*T*) and identifiers (*ID*) are identical to those in Table 1. The best values per column are printed in bold font.

T	*ID*	Area-based (%)			Object-based (%)			Object-based (50 m^2) (%)			*RMS* (m)
		$Comp_{ar}$	$Corr_{ar}$	Q_{ar}	$Comp_{obj}$	$Corr_{obj}$	Q_{obj}	$Comp_{50}$	$Corr_{50}$	Q_{50}	
P	HANC1	57.1	73.1	47.0	38.2	67.2	33.2	71.5	86.0	66.2	1.40
	HANC2	67.4	64.8	49.1	60.7	54.8	40.9	71.6	75.8	59.2	1.47
I	DLR	58.8	**76.6**	49.8	47.2	**75.1**	41.0	70.0	87.1	63.2	**1.30**
DO	CAL1	70.8	66.5	48.0	67.7	42.6	31.5	83.3	86.7	70.1	1.43
	CAL2	67.2	70.4	52.8	57.2	75.0	48.4	82.6	82.9	73.3	1.33
	LJU1	**75.0**	59.9	49.4	**71.0**	47.2	39.6	90.6	76.2	70.6	1.47
	LJU2	63.8	67.2	48.3	43.7	66.4	36.2	81.3	86.9	72.8	1.43
	TEH	56.9	51.3	36.9	50.5	20.4	17.0	80.8	52.0	46.3	1.60
	KNTU	74.3	63.5	52.1	66.5	68.0	**49.9**	92.9	74.5	70.5	1.50
	TUM	70.3	**76.6**	**57.8**	59.2	72.4	48.0	**93.2**	**98.3**	**91.6**	1.37
	WHUZ	52.8	67.4	42.2	41.8	57.5	31.9	63.5	84.5	56.9	1.53
PI	ITCM	49.2	69.4	40.4	37.4	65.3	31.2	72.5	87.9	65.8	1.50
	ITCR	64.0	66.9	47.9	51.2	65.4	41.0	85.8	87.5	75.4	1.53
Average areas 1–3 (422 trees)											


(a) LJU1

(b) LJU1

(c) HANC1

Fig. 5. Typical errors in tree detection causing problems for many tree detection methods (also other ones than shown in the figure): (a) Area 1: a row of trees situated on a step edge of the terrain next to a building. (b) Area 2: hedges erroneously classified as trees. (c) Area 3: missed small trees.

mal is not well defined) could not be assigned to any of the dominant roof planes. The rather irregular shapes of the boundary polygons show that in this case model generation is actually incomplete, because the intersections between neighbouring roof planes are probably not dealt with yet. Smooth tree canopies in the ALS data can also cause segmentation errors like an isolated false positive roof plane and the wrong extension of one of the dominant roof plane in Fig. 6d.

Matching errors in the DSM are typical cases leading to erroneous 3D reconstructions for purely image-based methods. This can be seen in Fig. 6e: rather than correctly separating the two dominant roof planes (corresponding to a saddleback roof), the right half of the building is reconstructed as one single plane in combination with a large dormer. Note that this is a result of the only fully-automatic method solely relying on images (FIE).

Table 9
Average quality metrics and their standard deviations as a function of the original input data used for tree detection (areas 1–3). *P*: ALS points; *I*: images only; *DO*: DSM grid from ALS and orthophoto; *PI*: original ALS points and images.

	P (2)	*I* (1)	*DO* (8)	*PI* (2)
$Comp_{ar}$ (%)	62.2 ± 7.3	58.8 ± 0.0	66.4 ± 8.0	56.6 ± 10.5
$Corr_{ar}$ (%)	69.0 ± 5.9	76.6 ± 0.0	65.3 ± 7.5	68.2 ± 1.8
Q_{ar} (%)	48.1 ± 1.5	49.8 ± 0.0	48.5 ± 6.6	44.1 ± 5.3
$Comp_{obj}$ (%)	49.4 ± 15.9	47.2 ± 0.0	57.2 ± 11.0	44.3 ± 9.8
$Corr_{obj}$ (%)	61.0 ± 8.8	75.1 ± 0.0	56.2 ± 18.5	65.3 ± 0.1
Q_{obj} (%)	37.1 ± 5.4	41.0 ± 0.0	37.8 ± 11.2	36.1 ± 6.9
$Comp_{50}$ (%)	71.6 ± 0.1	70.0 ± 0.0	83.5 ± 9.6	79.1 ± 9.4
$Corr_{50}$ (%)	80.9 ± 7.2	87.1 ± 0.0	80.3 ± 13.5	87.7 ± 0.3
Q_{50} (%)	62.7 ± 4.9	63.2 ± 0.0	69.0 ± 13.2	70.6 ± 6.8
RMS (m)	1.43 ± 0.05	1.30 ± 0.0	1.46 ± 0.09	1.52 ± 0.02

Interestingly the geometric errors, especially the planimetric ones, are in similar ranges in areas 1 and 3, despite the more com-

Table 10
Average quality metrics and their standard deviations as a function of the processing strategies used for tree detection (areas 1–3). The numbers in parenthesis are the number of submissions per group.

	Supervised (*SUP*) (6)	*Model-based* (*MOD*) (7)
$Comp_{ar}$ (%)	61.8 ± 10.2	65.2 ± 6.4
$Corr_{ar}$ (%)	69.1 ± 5.0	65.5 ± 8.0
Q_{ar} (%)	48.1 ± 6.4	47.6 ± 5.0
$Comp_{obj}$ (%)	50.6 ± 12.9	55.5 ± 10.4
$Corr_{obj}$ (%)	64.2 ± 6.7	56.0 ± 20.2
Q_{obj} (%)	39.2 ± 8.3	36.4 ± 10.0
$Comp_{50}$ (%)	77.5 ± 12.4	82.1 ± 6.3
$Corr_{50}$ (%)	84.5 ± 8.7	79.9 ± 13.0
Q_{50} (%)	68.3 ± 12.4	67.4 ± 10.1
RMS (m)	1.46 ± 0.06	1.44 ± 0.10

Table 11
Average quality metrics and their standard deviations as a function of the primitives that are classified in tree detection (areas 1–3). *P*: ALS points or (DSM or image) pixels; *S*: Segments. The numbers in parenthesis are the number of submissions per group.

	P (5)	*S* (8)
$Comp_{ar}$ (%)	64.4 ± 9.1	63.2 ± 8.2
$Corr_{ar}$ (%)	69.1 ± 5.6	66.0 ± 7.5
Q_{ar} (%)	49.6 ± 5.8	46.7 ± 5.3
$Comp_{obj}$ (%)	53.3 ± 12.5	53.3 ± 11.5
$Corr_{obj}$ (%)	64.0 ± 7.5	57.2 ± 19.0
Q_{obj} (%)	40.8 ± 8.2	35.7 ± 9.4
$Comp_{50}$ (%)	78.5 ± 13.6	80.9 ± 6.7
$Corr_{50}$ (%)	83.8 ± 9.5	80.9 ± 12.3
Q_{50} (%)	68.9 ± 13.8	67.2 ± 9.4
RMS (m)	1.45 ± 0.07	1.45 ± 0.10

plex shapes in area 1. However, due to the relatively large point spacing of ALS data (approximately 40 cm), uncertainties of the methods delivering the outlines may not become visible in these statistics.

Table 13 compares the average evaluation results achieved for Vaihingen for the ten ALS-only methods and the two purely image-based methods. Because of the imbalance in numbers between the ALS-based and the image-based contributions one should be careful with a comparison. It cannot be clearly decided whether performance gaps are caused by different properties of particular methods or by properties of different data sources, but nevertheless some trends can be observed. The two purely image-based methods CKU and FIE do not outperform methods using ALS in terms of planimetric accuracy although the GSD of the images is much smaller. As already mentioned for building detection, this may be a hint that image-based methods presented here do not fully exploit all information contained in these very high-resolution images yet. However, the completeness of reconstructed roof planes is better for the image-based method, which could be explained by the relatively small ALS point density. Although quite different modelling strategies are applied, the topologic errors do not differ much, at least for methods using ALS data. This shows that in general segmentation of planes in ALS data works reliably today within the limitations discussed above.

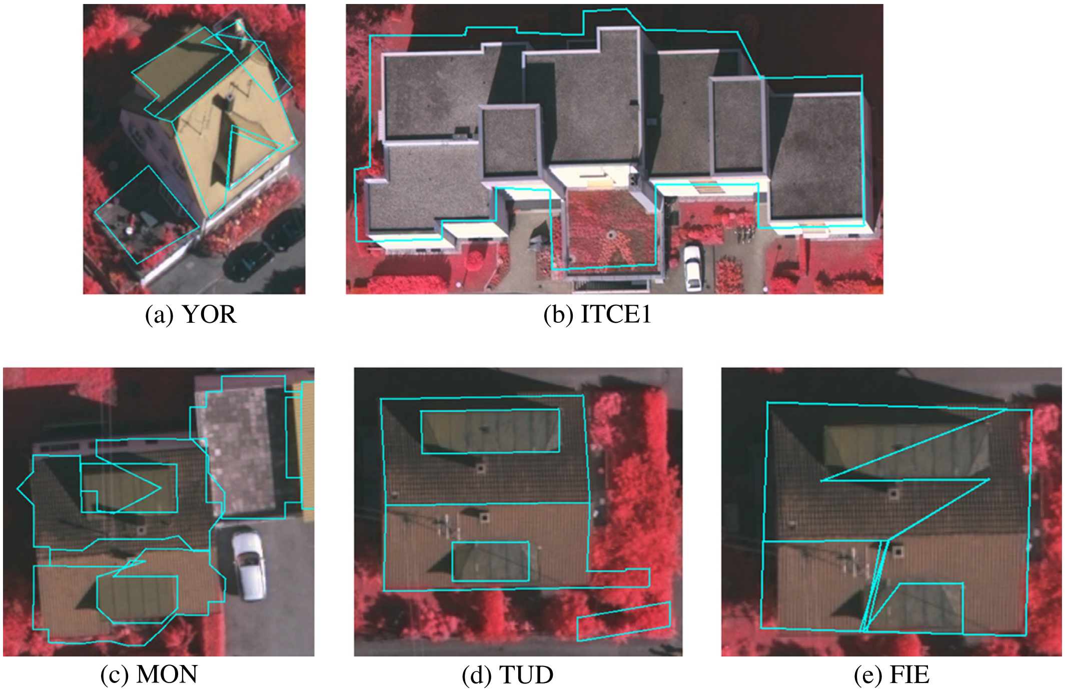
From a practical application viewpoint a larger correctness is considered more important than a larger completeness (Mayer et al., 2006), and we can generally observe this trend here. All methods are tuned to achieve a high correctness rather than a high completeness. Another interesting finding concerning practical applications is that semi-automated methods do not really outperform fully automatic ones, see Table 14 for a comparison.

5.2.1.2. Toronto: results. For areas 4 and 5 we only received three submissions (YOR, CKU, and FIE). The average evaluation results are shown in Table 15. Like area 2 in Vaihingen, both Toronto areas mostly contain flat roofs, but with a larger height variation and shape complexity. For all methods, completeness and correctness are worse than in Vaihingen. Under-segmentation but also clusters where both, under- and over-segmentation occur (*N*:*M*) are the dominant errors evoked by the strongly varying and rarely symmetric roof shapes. Naturally, this results in higher RMS values, too. The relatively high RMSZ value of YOR (7.95 m) reflects partially erroneous segmentation of roofs that predominantly occurs at high-rise buildings.

In Toronto the methods face the problem of large flat roofs with many superstructures and plane variations. While the methods CKU, FIE and YOR produce hardly any topologic *N*:*M* errors (i.e., clusters of planes of combined over- and under-segmentation) in Vaihingen, this kind of error is quite significant in Toronto, where simple 1:*M* and *N*:1 errors also occur more frequently than in Vaih-

Table 12
Evaluation of the building reconstruction results: average of areas 1–3. The identifiers (*ID*) are identical to those in Table 2. The remaining column headings are explained in Section 3.2. The identifiers printed in bold font indicate methods for which we did not receive results for all areas; in this case, the numbers in parenthesis indicate the areas for which results were received. The best values per column are printed in bold font. To have an unbiased ranking, the methods only contributing to one area (FIE, BNU, KNTU) are not considered for the ranking.

	Name	Per-roof plane (%)			Per-roof plane (10 m^2) (%)			Topology			*RMSXY* (m)	*RMSZ* (m)
		$Comp_{obj}$	$Corr_{obj}$	Q_{obj}	$Comp_{10}$	$Corr_{10}$	Q_{10}	$N_{1:M}$	$N_{N:1}$	$N_{N:M}$		
P	MON	77.5	89.7	71.2	90.3	91.4	83.5	6	30	3	0.90	0.37
	VSK	74.2	98.6	73.5	86.1	98.6	85.2	4	32	2	0.83	0.27
	ITCE1	69.4	90.1	63.1	78.4	90.3	69.5	5	27	6	1.00	**0.17**
	ITCE2	69.8	98.3	68.7	76.8	**100.0**	76.8	**0**	30	1	1.03	**0.17**
	ITCX1	69.5	98.1	68.7	74.4	98.0	73.2	2	31	1	0.70	0.20
	ITCX2	82.0	92.9	76.8	91.0	98.1	89.3	11	31	3	0.70	0.27
	ITCX3	**82.8**	94.9	78.7	**93.2**	97.8	91.2	4	31	3	0.70	0.20
	CAS (2,3)	68.5	**100.0**	68.5	81.2	**100.0**	81.2	2	**26**	**0**	0.75	0.25
	TUD	70.0	95.8	67.8	78.8	98.6	78.0	2	**26**	1	0.70	0.20
	YOR	79.9	99.5	79.5	91.8	99.7	**91.6**	4	30	5	**0.63**	0.27
D	**KNTU** (3)	80.4	96.7	78.3	91.9	97.7	90.0	0	52	0	0.90	0.40
I	**FIE** (3)	82.6	83.1	70.7	88.7	93.4	83.5	7	44	5	1.10	0.40
	CKU	82.1	96.8	**80.1**	91.4	99.4	90.9	7	29	2	0.73	0.63
PI	**BNU** (3)	87.2	100.0	87.2	96.0	100.0	97.1	2	52	0	0.60	0.10
Average areas 1–3 (592 roof planes)												



(a) YOR (b) ITCE1

(c) MON (d) TUD (e) FIE

Fig. 6. Examples for common problems in 3D building reconstruction. (a) Missing small roof planes (e.g. sides of the dormer having a triangular roof plane) and missing small appendices to bigger roof planes (e.g. below the dormer having a triangular roof plane and to the left of the larger rectangular dormer). (b) Under-segmentation of complex flat roof building. (c) Over-segmentation and incomplete model generation. (d) and (e) Wrong segmentations, in case (e) probably due to matching errors.

ingen. This observation, in conjunction with the occurrence of large height errors, confirms that complex flat roof structures as occurring in typical CBD-like areas are still a challenge for automatic reconstruction.

Table 13
Average quality metrics and their standard deviations as a function of the original input data used for building reconstruction (areas 1–3). *P*: ALS points; *I*: aerial images. The numbers in parenthesis are the number of submissions per group. Input data used only by one method (*D*, *PI*; cf. Table 1 and related text) are omitted.

	P (10)	*I* (2)
$Comp_{obj}$ (%)	74.4 ± 5.7	82.4 ± 0.3
$Corr_{obj}$ (%)	95.8 ± 3.8	90.0 ± 9.7
Q_{obj} (%)	71.6 ± 5.3	75.4 ± 6.6
$Comp_{10}$ (%)	84.2 ± 7.1	90.0 ± 1.9
$Corr_{10}$ (%)	97.3 ± 3.5	96.4 ± 4.3
Q_{10} (%)	82.0 ± 7.6	87.2 ± 5.3
$N_{1:M}$	4 ± 3	7 ± 0
$N_{N:1}$	29 ± 2	36 ± 10
$N_{N:M}$	3 ± 2	3 ± 2
RMSXY (m)	0.8 ± 0.1	0.9 ± 0.3
RMSZ (m)	0.2 ± 0.1	0.5 ± 0.2

Table 14
Average quality metrics and their standard deviations as a function of the degree of automation of the methods used for building reconstruction (areas 1–3). *S*: semi-automatic methods; *F*: fully-automatic methods. The numbers in parenthesis are the number of submissions per group.

	S (2)	*F* (12)
$Comp_{obj}$ (%)	75.8 ± 8.9	77.0 ± 6.4
$Corr_{obj}$ (%)	96.3 ± 2.5	95.0 ± 5.2
Q_{obj} (%)	73.8 ± 1.7	73.7 ± 6.6
$Comp_{10}$ (%)	78.1 ± 5.9	87.0 ± 6.6
$Corr_{10}$ (%)	96.5 ± 3.7	97.1 ± 3.5
Q_{10} (%)	75.8 ± 2.9	84.7 ± 7.7
$N_{1:M}$	5 ± 3	4 ± 3
$N_{N:1}$	30 ± 3	34 ± 10
$N_{N:M}$	1 ± 1	2 ± 2
RMSXY (m)	0.7 ± 0.0	0.8 ± 0.2
RMSZ (m)	0.4 ± 0.3	0.3 ± 0.1

5.2.1.3. Discussion. Comparing the five test areas, it seems that area 2 in Vaihingen offers the most favourable conditions for automatic roof reconstruction because medium-size flat roof buildings dominant here. In all other areas, complex roof structures cause a considerable amount of segmentation errors. In general, the main roof structures are represented well (and certainly good enough for visualisation) if the basic roof shape is relatively simple and if there are no dormers or only dormers that are small compared to the dominant roof planes. Otherwise, the algorithms compared in this test frequently produce incorrect and inaccurate results. This fact and an analysis of the geometrical errors show that methods for roof plane reconstruction still have room for improvement, independently from the data source. On average, all approaches satisfy the standards required for practical relevance according to Mayer et al. (2006), at least if focus is on large roof planes.

For future research the focus can be on smaller roof structures and a better treatment of step edges in complex flat roof buildings. The first issue is also linked to the density of point clouds in case ALS data are used; see the discussion on small objects in Section 5.1. Although image based point clouds in general can offer a better point density, this potential has not been fully exploited, yet, by the corresponding methods. There still remains room for improvements. Concerning the modelling of complex flat roof structures, like in CBD areas, the point cloud density also plays a major role. With the advent of denser point clouds, also from ALS, such problems might be mitigated.

6. Conclusion

In this paper, several methods from current research in urban object extraction were compared based on a benchmark data set.

Table 15
Evaluation of the building reconstruction results: average of areas 4 and 5. The identifiers (*ID*) are identical to those in Table 2. The remaining column headings are explained in Section 3.2. The best values per column are printed in bold font. FIE is not considered in the ranking because it only contributes to area 4.

	Name	Per-roof plane (%)			Per-roof plane (10 m^2) (%)			Topology			*RMSXY* (m)	*RMSZ* (m)
		$Comp_{obj}$	$Corr_{obj}$	Q_{obj}	$Comp_{10}$	$Corr_{10}$	Q_{10}	$N_{1:M}$	$N_{N:1}$	$N_{N:M}$		
P	**YOR**	**70,0**	**91,7**	**66,2**	**86,4**	**92,1**	**80,4**	**16**	84	**22**	**0,90**	7,95
I	**CKU**	69,5	81,8	60,1	79,1	81,4	67,1	27	**60**	64	1,75	**4,80**
	FIE	52,3	91,5	49,9	60,4	91,9	57,3	56	62	36	1,40	2,70
Average areas 4–5 (1607 roof planes)												

The results achieved by the methods for building detection show that this task can be satisfactorily solved for buildings larger than 50 m^2 by methods relying on different processing strategies and different sensor data, but there is still room for improvement in detecting small building structures and in precise delineation of the building boundaries. Most of the methods for tree detection were successful in detecting large trees under favourable conditions, but failed to do so in very complex inner city environments. Small trees could not be detected reliably by any of the methods, either; this seems to indicate a field requiring further research. The results achieved for 3D building reconstruction show the potential, but also the limitations of state-of-the-art methods. While the problem may be considered to be solved for visualisation purposes, the production of geometrically and topologically correct LoD2 building models still poses challenges in difficult urban environments. In particular, no method seems to be able to fully exploit the accuracy potential inherent in the sensor data. It would be desirable to receive more results solely based on images to obtain a more realistic assessment of the potential inherent in that data source.

The test data sets are still available, and results are continuously received and evaluated. Currently, we have only a rather small number of submissions solely relying on images, and we would like to attract more participants using these data to obtain a more general view on the potential of this data source for urban object extraction. For methods that are not specifically tailored to using a specific data source, but can cope with a generic set of features, it would be interesting to receive results achieved for different input data sets, in order to make a comparison of the potential of different data sets more conclusive. In the future, we also want to expand the benchmark. A reference for image labelling that will also include training data for supervised methods is under preparation. Furthermore, we have received a third test data set, for which the reference is under preparation. It is the goal of these efforts to provide a benchmark data set as a basis for making current and future developments in urban object extraction more comparable.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments which certainly helped to improve this paper. The Vaihingen data set was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF; Cramer, 2010): http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html. The reference for Vaihingen was generated by RAG Steinkohle AG and SIRADEL (http://www.siradel.com). The authors would like to acknowledge the provision of the Toronto data set by Optech Inc. (http://www.optech.ca), First Base Solutions Inc. (http://www.firstbasesolutions.com) and York University (http://geoict.yorku.ca). The reference for Toronto was created by York University. The ISPRS benchmark is available via the WWW site of ISPRS Working Group III/4, "3D Scene Analysis": (http://www2.isprs.org/commissions/comm3/wg4.html).

References

Awrangjeb, M., Zhang, C., Fraser, C.S., 2012. Building detection in complex scenes through effective separation of buildings from trees. Photogrammetric Engineering & Remote Sensing 78 (7), 729–745.

Bulatov, D., Häufel, G., Meidow, J., Pohl, M., Solbrig, P., Wernerus, P., 2014. Context-based automatic reconstruction and texturing of 3D urban terrain for quick-response tasks. ISPRS Journal of Photogrammetry and Remote Sensing 93, 157–170.

Champion, N., Rottensteiner, F., Matikainen, L., Liang, J., Hyyppä, J., Olsen, B., 2009. A test of automatic building change detection approaches. International Archives of Photogrammetry, Remote Sensing and Spatial Information Systems 38 (Part 3-W4), 145–150.

Cramer, M., 2010. The DGPF test on digital aerial camera evaluation – overview and test design. Photogrammetrie – Fernerkundung – Geoinformation 2, 73–82.

Dorninger, P., Pfeifer, N., 2008. A comprehensive automated 3D approach for building extraction, reconstruction, and regularization from airborne laser scanning point clouds. Sensors 8 (11), 7323–7343.

Everingham, M., van Gool, L., Williams, C., Winn, J., Zisserman, A., 2010. The Pascal Visual Object Classes (VOC) challenge. International Journal of Computer Vision 88 (2), 303–338.

Gerke, M., Xiao, J., 2014. Fusion of airborne laserscanning point clouds and images for supervised and unsupervised scene classification. ISPRS Journal of Photogrammetry and Remote Sensing 87, 78–92.

Grigillo, D., Kanjir, U., 2012. Urban object extraction from digital surface model and digital aerial images. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences I-3, 215–220.

Gröger, G., Kolbe, T.H., Czerwinski, A., Nagel, C., 2008. OpenGIS city geography markup language (CityGML) encoding standard, Version 1.0.0, OGC Doc. No. 08-007r[1] <http://www.opengeospatial.org/standards/citygml> (10/01/12).

ISPRS, 2013. Web site of the ISPRS test project on urban classification and 3D building reconstruction <http://www2.isprs.org/commissions/comm3/wg4/tests.html> (20/08/13).

Kaartinen, H., Hyyppä, J., Gülch, E., Hyyppä, H., Matikainen, L., Hofmann, A.D., Mäder, U., Persson, A., Söderman, U., Elmqvist, M., Ruiz, A., Dragoja, M., Flamanc, D., Maillet, G., Kersten, T., Carl, J., Hau, R., Wild, E., Frederiksen, L., Homgaard, J., Vester, K., 2005. Accuracy of 3D city models: EuroSDR comparison. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 36 (Part 3-W19), 227–232.

Liu, C., Shi, B., Xuan, X., Nan, L., 2012. LEGION segmentation for building extraction from LiDAR data. International Archives of Photogrammetry, Remote Sensing and Spatial Information Systems 39 (Part 3), 291–296.

Mayer, H., 2008. Object extraction in photogrammetric computer vision. ISPRS Journal of Photogrammetry and Remote Sensing 63 (2), 213–222.

Mayer, H., Hinz, S., Bacher, U., Baltsavias, E., 2006. A test of automatic road extraction approaches. International Archives of Photogrammetry, Remote Sensing and Spatial Information Systems 36 (Part 3), 209–214.

Meidow, J., Schuster, H.-F., 2005. Voxel-based quality evaluation of photogrammetric building acquisitions. International Archives of the photogrammetry, remote sensing and spatial information sciences XXXVI (Part 3/W24), 117–122.

Mongus, D., Lukač, N., Obrul, D., Žalik, B., 2013. Detection of planar points for building extraction from LiDAR data based on differential morphological and attribute profiles. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences II-3/W1, 21–26.

Mongus, D., Lukač, N., Žalik, B., 2014. Ground and building extraction from LiDAR data based on differential morphological profiles and locally fitted surfaces. ISPRS Journal of Photogrammetry and Remote Sensing 93, 145–156.

Moussa, A., El-Sheimy, N., 2012. A new object based method for automated extraction of urban objects from airborne sensors data. International Archives of Photogrammetry, Remote Sensing and Spatial Information Systems 39 (Part 3), 309–314.

Niemeyer, J., Wegner, J.D., Mallet, C., Rottensteiner, F., Soergel, U., 2011. Conditional random fields for urban scene classification with full waveform LiDAR data. In: Stilla, U., et al. (Eds.), PIA 2011, LNCS, vol. 6952, pp. 233–244.

Niemeyer, J., Rottensteiner, F., Soergel, U., 2013. Classification of urban LiDAR data using conditional random field and random forests. In: Proceedings of the Joint Urban Remote Sensing Event (JURSE), São Paulo, Brazil, pp. 139–142.

Oude Elberink, S., Vosselman, G., 2009. Building reconstruction by target based graph matching on incomplete laser data: analysis and limitations. Sensors 9 (8), 6101–6118.

Oude Elberink, S., Vosselman, G., 2011. Quality analysis on 3D building models reconstructed from airborne laser scanning data. ISPRS Journal of Photogrammetry and Remote Sensing 66 (2), 157–165.

Perera, S., Mass, H.-G., 2014. Cycle graph analysis for 3D roof structure modelling: concepts and performance. ISPRS Journal of Photogrammetry and Remote Sensing 93, 213–226.

Rau, J.-Y., Lin, B.-C., 2011. Automatic roof model reconstruction from ALS data and 2D ground plans based on side projection and the TMR algorithm. ISPRS Journal of Photogrammetry and Remote Sensing 66 (6), s13–s27 (supplement).

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.

Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., Breitkopf, U., 2012. The isprs benchmark on urban object classification and 3d building reconstruction. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences I-3, 293–298.

Rutzinger, M., Rottensteiner, F., Pfeifer, N., 2009. A comparison of evaluation techniques for building extraction from airborne laser scanning. IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sens. 2 (1), 11–20.

Scharstein, D., Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision 47 (1/2/3), 7–42.

Schindler, K., Förstner, W., Paparoditis, N., 2011. Proceedings of the 1st IEEE/ISPRS Workshop on Computer Vision for Remote Sensing of the Environment, 2011 IEEE International Conference on Computer Vision Workshops, 6–13 November 2011, Barcelona, Spain, available via IEEE Explore.

Sohn, G., Huang, X., Tao, V., 2008. Using binary space partitioning tree for reconstructing 3D building models from airborne LiDAR data. Photogrammetric Engineering & Remote Sensing 74 (11), 1425–1440.

Sohn, G., Jwa, Y., Kim, H.B., Jung, J., 2012. An implicit regularization for 3D building rooftop modelling using airborne LiDAR data. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 1 (3), 305–310.

Sohn, G., Rottensteiner, F., Wegner, J.D., (Eds.), 2013. Proceedings of the ISPRS Workshop on 3D Virtual City Modeling (VCM 2013), International Annals of the Photogrammetry, Remote Sensing and Spatial, Information Sciences, vol. II-3/W1.

Stilla, U., Rottensteiner, F., Mayer, H., Jutzi, B., Butenuth, M. (Eds.), 2011. Proceedings of the ISPRS Conference on Photogrammetric Image Analysis 2011 (PIA 2011). Lecture Notes in Computer Sciences (LNCS), vol. 6952. Springer, Heidelberg.

Wei, Y., Yao, W., Wu, J., Schmitt, M., Stilla, U., 2012. Adaboost-based feature relevance assessment in fusing LiDAR and image data for classification of trees and vehicles in urban scenes. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences I-7, 323–328.

Xiong, B., Oude Elberink, S., Vosselman, G., 2014. A graph edit dictionary for correcting errors in roof topology graphs reconstructed from point clouds. ISPRS Journal of Photogrammetry and Remote Sensing 93, 227–242.

Yang, M.Y., Förstner, W., 2011. A hierarchical conditional random field model for labelling and classifying images of man-made scenes. In: IEEE International Conference on Computer Vision Workshops (WS04), pp. 196–203.

Yang, B.S., Xu, W.X., Dong, Z., 2013. Automated building outlines extraction from airborne laser scanning point clouds. IEEE Geoscience and Remote Sensing Letters, 5 pages, doi: 10.1109/LGRS.2013.2258887 (in press).

Zhan, Q., Liang, Y., Wei, C., Xiaoa, Y., 2012. Ground object recognition using combined high resolution airborne images and DSM. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XXXIX-B3, 573–577.

Zhang, W., Grussenmeyer, P., Yan, G., Mohamed, M., 2011. Primitive-based building reconstruction by integration of Lidar data and optical imagery. International Archives of Photogrammetry, Remote Sensing and Spatial Information Systems 38 (Part 5-W12).