THE ISPRS BENCHMARK ON URBAN OBJECT CLASSIFICATION AND 3D BUILDING RECONSTRUCTION

F. Rottensteiner [a, *], G. Sohn [b], J. Jung [b], M. Gerke [c], C. Baillard [d], S. Benitez [d], U. Breitkopf [a]

[a] Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Germany – (rottensteiner, breitkopf)@ipi.uni-hannover.de
[b] GeoICT Lab; Earth and Space Science and Engineering Department, York University, Toronto, Canada – (gsohn, jwjung)@yorku.ca
[c] University of Twente, Faculty ITC; EOS department, Enschede, The Netherlands – gerke@itc.nl
[d] SIRADEL, Rennes, France – (cbaillard, sbenitez)@siradel.com

Commission III, WG III/4

KEY WORDS: Automatic object extraction, 3D building reconstruction, aerial imagery, laser scanning, evaluation, 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. Researchers were encouraged to submit 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 at least 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. The success of the Middlebury Stereo Vision test (Scharstein & Szeliski, 2002) has shown the importance of providing common data sets with ground truth for comparing different approaches to problems in computer vision. Such a comparison 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. There have been attempts in the past to distribute data sets for benchmarking object extraction methods, e.g. the OEEPE/EuroSDR data sets for building (Kaartinen et al., 2005) and road extraction (Mayer et al., 2006) and for automated updating of maps (Champion et al., 2009). As far as data from aerial sensors are concerned, these data sets are outdated; 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 airborne laserscanner (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, 2012). Unlike previous benchmark data sets on urban object detection, the reference data included 2D outlines of multiple object types. It also contains different types of urban development. Researchers were given access to the sensor data and encouraged to carry out one or more of several urban object extraction tasks. The goal of **Task 1, urban object detection**, was to determine 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. The goal of **Task 2, 3D building reconstruction**, was to reconstruct 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. This paper gives a report about the evaluation of the results submitted by the test participants. As far as object detection is concerned, it is restricted to the object classes most frequently submitted by the participants, namely buildings and trees.

2. DATA AND TEST SETUP

Data Set 1 – Vaihingen (Germany): This is a subset of the data used for the test of digital aerial cameras carried out by the German Association of Photogrammetry, Remote Sensing, and Geoinformation (DGPF) (Cramer, 2010). It consists of 20 images of the high-resolution DMC block and subsets of 5 of the ALS strips used in that test. The images are 16 bit pan-sharpened colour infrared (CIR) images with a ground sampling distance (GSD) of 8 cm (flying height above ground: 800 m, focal length: 120 mm, 65% forward lap, 60% side lap), taken with an Intergraph/ZI DMC. A fourfold overlap is ensured for the entire test site. Orientation data are distributed with the images (georeferencing accuracy: 1 pixel). The ALS data were acquired using a Leica ALS50 system with 45° field of view and a mean flying height above ground of 500 m. The average strip overlap is 30% and the point density varies between 4 and 7 points/m^2. Multiple pulses were recorded. The point cloud

* Corresponding author.

was pre-processed to compensate for systematic offsets between the strips and between the ALS and image data. 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: **Area 1** is characterized by dense development consisting of historic buildings having rather complex shapes, roads and trees. **Area 2** is characterized by a few high-rising residential buildings that are surrounded by trees. **Area 3** 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. The planimetric accuracy of well-defined corner points is 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.

Data Set 2 – Toronto (Canada): This data set also consists of image and ALS data. There are 13 RGB colour images (8 bit) taken with a Microsoft Vexcel UltraCam-D and having a GSD of 15 cm. The images are arranged in a block of 3 strips (flying height above ground: 1600 m, focal length: 101.4 mm, 60% forward lap, 30% side lap). Orientation data were also distributed with the images (georeferencing accuracy: 1 pixel). Optech's ALTM-ORION M was used to acquire the ALS data at a flying height of 650 m. The sensor operates at a wavelength of 1064 nm and scans the underlying topography with a scan width of 20°. The data set consists of 6 strips and the point density is 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.

Two test sites were selected: **Area 4** 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** represents a cluster of high-rise buildings typical for American cities. The scene contains shadows cast by high buildings and various types of urban objects. The reference for building reconstruction (LoD2) was generated by stereo plotting. The accuracy of well-defined points is 20 cm in planimetry and 15 cm in height.

Task 1- Urban object detection: The goal of the first task was the detection of objects in the test areas. The participants could deliver outline polygons of the objects or binary object masks. Results could be submitted for any of the object types for which reference data were available, but most of the participants only submitted results for buildings and trees.

The results submitted by the participants were compared to the reference data. For the evaluation of the thematic accuracy, the method described in (Rutzinger et al., 2009) was used. After a topological clarification, the *completeness* and the *correctness* of the results were determined on a per-area level (Cm_{ar} / Cr_{ar}) and on a per-object level (Cm_{ob} / Cr_{ob}). We also report per-object completeness only for objects larger than 50 m^2 (Cm_{50} / Cr_{50}), corresponding to the most relevant buildings and to trees having crown diameters larger than 8.4 m, to analyse the dependency of per-object quality metrics on the object size. For the object-based metrics, we required a minimum overlap of 50% for an object with the reference to be counted as a true positive. We also evaluated the geometrical quality of the detected objects (*RMS*). For buildings, this is the RMS error of the planimetric distances of the extracted boundary points to their nearest neighbours on the corresponding reference boundaries; for trees, it is the RMS error of the planimetric distances between the centres of gravity of corresponding objects. Only distances shorter than 3 m are considered.

Task 2- Building reconstruction: The goal of the second task was 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 focused on an analysis of the segmentation quality and on the geometrical errors of the submitted models. The analysis of the quality of the segmentation was based on a comparison of roof plane label images carried out similarly to the overlap analysis for the evaluation of object detection, but without topological clarification. The completeness and the correctness of the extracted roof planes are reported on a per-plane basis (Cm_{ob} / Cr_{ob}). These numbers refer to the number of roof planes in one data set having at least 50% overlap with planes in the other data set. Per-roof-plane completeness and correctness are also reported for planes covering an area of at least 10 m^2 (Cm_{10} / Cr_{10}), again to analyse the dependency of these indicators from the object size. The correspondence analysis provides the numbers of instances where *1:M*, *N:1*, and *N:M* relations between roof planes in the reference and planes in the reconstruction results occur ($N_{1:M}$ / $N_{N:1}$ / $N_{N:M}$). $N_{1:M}$ is an indicator for oversegmentation, $N_{N:1}$ for undersegmentation. $N_{N:M}$ indicates clusters of planes that are both over- and undersegmented. These numbers also reflect the quality of the roof plane segmentation. 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 extracted roof plane boundary points to their nearest neighbours on the corresponding reference boundaries (*RMS*). The RMS errors of the height differences *RMSZ* were 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.

3. METHODS

3.1 Task 1: Urban Object Detection

For the urban object detection task, results were submitted for eight different methods. Five are mainly rule-/knowledge based, whereas three pursue a supervised classification methodology.

A. Moussa, University of Calgary, Canada (CAL): In this approach, the ALS point cloud is used in combination with an ortho-rectified CIR image. The method starts with a rule-based segmentation and classification of the ALS data into building, tree and ground segments. Spectral information obtained from the image is used to refine the classification, and morphologic operations are applied to smooth the resulting label image (Moussa & El-Sheimy, 2012).

D. Bulatov, Fraunhofer Institute Ettlingen, Germany (FIE): Here only the images are used. After image matching and the generation of a digital terrain model (DTM), a rule-based classification of buildings and vegetation is carried out, and the boundary polygons of the buildings are extracted. Optionally, the polygons can be regularised (Bulatov et al., 2011).

J. Niemeyer, University of Hannover, Germany (HAN): This approach uses a supervised classification of the ALS points based on Conditional Random Fields, which incorporates a

statistical model of context (Niemeyer et al., 2011). The resulting point clusters of each class are projected into a label image, which finally is smoothed by morphological operations.

P. Dorninger, Vermessung Schmid, Klosterneuburg, Austria (VSK): This approach solely relies on the ALS point cloud to detect buildings. After detecting planar surface patches in the point cloud, the detection result is refined by a model-based classification and by the combination of patches. The geometry is enhanced using morphological operations. Finally, the borders of the regions are delineated, which results in polygonal building outlines (Dorninger & Pfeifer, 2008).

D. Grigillo and U. Kanjir, University of Ljubljana, Slovenia (LJU): The ALS data are used to derive a DSM, a DTM and a normalised DSM (nDSM) of the area. Using the nDSM, a mask of objects above ground is computed, and vegetation is separated from buildings using the NDVI derived from a CIR orthoimage. The remaining regions are morphologically filtered, and the outline of buildings is approximated by a method based on Hough transformation (Grigillo et al. 2011).

Q. Zhan, Wuhan University, China (WHU): After ortho-rectifying the images using the ALS DTM, colour and height information is used to detect the different land cover classes by a supervised classification approach. Then, spectrally similar classes like building/road and tree/low vegetation are separated using a method based on an analysis of the elevation histogram.

C. Liu, Tongji University, China (TON): A LEGION (locally excitatory globally inhibitory oscillator network) segmentation (Liu & Wang, 1999) is applied to the ALS point cloud. After that, texture features are used in a classifier based on neuronal networks. All the parameters for the method are pre-set by the authors. The outlines of the clusters are refined and regularized in order to have rectangular outlines using a least squares estimation method (Liu et al., 2012).

W. Yao, TU Munich, Germany (TUM): A supervised classification approach employing both height and image data is applied in this method. For each cell of a grid defined in object space, 6 colour and height features as well as 7 features encoding local context are determined. These features are fed into an AdaBoost classifier. In order to obtain training data, ground truth was digitized manually in 10% of each area.

3.2 Task 2: 3D Building Reconstruction

For this task, results obtained by seven different methods were submitted. Whereas five methods work fully automatically, two rely on some human intervention.

P. Dorninger, Vermessung Schmid, Klosterneuburg, Austria (VSK): Starting from the detected building outlines (cf. Section 3.1), wall hypotheses are generated. The planar segmentation already employed in building detection is used to find and combine roof planes (Dorninger & Pfeifer, 2008).

J.-Y. Rau, National Cheng-Kung University Taiwan (CKU): This approach uses the original images together with a building map. After manually measuring roof structure lines, a Delaunay triangulation of the roof points constrained on the 3D structure lines is performed. The triangles and, thus, the roof planes are refined and merged based on an analysis of the projection of the outlines into individual images (Rau & Lin, 2011).

S. Oude Elberink, University of Twente, The Netherlands (ITCE1/ITCE2): A coarse building map is used to identify building points in an ALS point cloud. The building points are segmented using a plane-based approach. An adjacency graph representing the roof plane topology is matched against a library of predefined roof primitives to eliminate wrong extractions and to find possibly missing planes. Two different results were submitted: ITCE1 is based on a more model-driven parameter setting, whereas ITCE2 was obtained by a more data-driven approach (Oude Elberink & Vosselman 2009, 2011).

B. Xiong, University of Twente, The Netherlands (ITCX): This is an extension of the approach by Oude Elberink (see above). Graph matching is carried out using an extended primitive library to identify general buildings, even buildings with highly complex structures. In an interactive step, a human operator is asked to identify data errors, e.g. false and missed intersection lines or roof segments. The algorithm uses this additional knowledge to update the library and tries to refine other buildings as well.

D. Bulatov, Fraunhofer Institute Ettlingen, Germany (FIE): After building detection (cf. Section 3.1) a detailed roof analysis is performed, using a RANSAC based approach for the initial clustering of normal vectors (Bulatov et al., 2011).

W. Zhang, Beijing Normal University, China (BNU): This is a model-based building reconstruction method using images and ALS points. The initial geometric parameters of each building roof are retrieved from the ALS point cloud. The selected building prototype is then re-projected into the images and its geometry is refined (Zhang et al., 2011).

G. Sohn, York University, Canada (YOR): This method is based on the integration of two approaches described in (Sohn et al., 2008) and (Jwa et al., 2008). After detecting buildings, a Binary Space Partitioning tree produces initial approximations of roof polygons. A regularization operator based on the Minimum Description Length principle removes topological errors by implicitly adapting the roof polygons for achieving maximal geometrical regularity.

4. RESULTS AND DISCUSSION

4.1 Task 1: Urban Object Extraction

The evaluation of the building and tree detection results is summarized in Tables 1 and 2, respectively. For each quality measure and area the best result is shown in bold font style. No tree extraction results were submitted for Toronto. The area-based quality metrics and the RMS errors for tree detection must be interpreted with caution because they are affected by the generalization errors of the reference.

Area 1: Except for TON and WHU, all methods for building detection achieve an area-based completeness between 85% and 93% and a correctness between 90% and 98%. The values for Cm_{ar} are lower than Cr_{ar} for most methods due to a large flat building part (missed by all methods except LJU). All methods perform much better for larger buildings than for smaller ones. CAL achieves $Cm_{50} = Cr_{50} = 100\%$. HAN confused some large trees with buildings, resulting in 12% false positive detections larger than 50 m^2. This is also one of the reasons for the poor performance of HAN in detecting trees. However, all tree detection methods have problems in this area; only TUM and LJU detect more than 50% of the trees, and only for TUM and CAL more than 50% of the detected trees are correct. The only methods achieving completeness and correctness higher than 80% for larger trees (Cm_{50} / Cr_{50}) are TUM and LJU. The main problem is a row of trees in the shadow of a multi-storey building and along a road having a lower elevation than the surrounding terrain. It is missed by all methods except TUM.

Area 2: Nearly all methods perform better in this area than in area 1. The per-building metrics show that except for TON and WHU, all methods can detect 100% of the buildings larger than

50 m^2. Except for WHU and HAN, all detected buildings are correct. LJU produces a few large false positives buildings in areas with terrain discontinuities. All methods except LJU achieve completeness and correctness values larger than 85% for trees larger than 50 m^2. TUM is the only method achieving 100% for both values. The per-object quality metrics for all trees are generally better than in area 1, but all approaches either deliver many false positives or many false negatives.

Name	Cm_{ar} / Cr_{ar} [%]	Cm_{ob} / Cr_{ob} [%]	Cm_{50} / Cr_{50} [%]	RMS_e [m]
Area 1 (37 buildings; 125 m x 200 m)				
CAL	89.1 / 94.7	83.8 / **100.0**	**100.0** / **100.0**	0.77
HAN	87.0 / 90.1	83.8 / 72.1	**100.0** / 87.9	1.09
LJU	**93.2** / 94.1	81.1 / **100.0**	96.7 / **100.0**	0.74
TON	76.7 / 95.7	75.7 / 93.5	93.3 / 96.7	1.18
TUM	89.8 / 90.1	**89.2** / 91.7	96.7 / 93.5	**0.71**
VSK	85.7 / **98.1**	78.4 / **100.0**	96.7 / **100.0**	0.99
WHU	84.4 / 83.9	78.4 / 43.5	86.7 / 96.3	1.01
Area 2 (14 buildings; 170 m x 190 m)				
CAL	93.2 / 95.4	78.6 / **100.0**	**100.0** / **100.0**	0.73
HAN	93.8 / 91.4	78.6 / 52.4	**100.0** / 84.6	0.71
LJU	**95.1** / 94.3	**85.7** / **100.0**	**100.0** / **100.0**	0.77
TON	88.5 / **98.9**	71.4 / **100.0**	90.9 / **100.0**	0.71
TUM	92.5 / 93.9	78.6 / **100.0**	**100.0** / **100.0**	**0.64**
VSK	85.4 / 98.4	**85.7** / **100.0**	**100.0** / **100.0**	1.17
WHU	79.6 / 91.9	57.1 / 42.3	72.7 / 90.9	0.87
Area 3 (56 buildings; 150 m x 220 m)				
CAL	87.0 / 95.2	66.1 / **100.0**	87.5 / **100.0**	0.54
FIE	89.0 / 86.9	78.6 / **100.0**	**97.5** / **100.0**	0.99
HAN	93.8 / 93.7	**82.1** / 90.2	**97.5** / **100.0**	0.65
LJU	**94.4** / 95.4	**82.1** / **100.0**	**97.5** / **100.0**	**0.52**
TON	67.8 / 98.4	55.4 / **100.0**	77.5 / **100.0**	1.17
TUM	86.8 / 92.5	75.0 / **100.0**	**97.5** / **100.0**	0.70
VSK	86.3 / **98.7**	75.0 / **100.0**	95.0 / **100.0**	0.81
WHU	76.9 / 92.6	64.3 / 79.2	77.5 / **100.0**	0.73
Area 4 (58 buildings; 530 m x 600 m)				
TUM	**85.1 / 80.0**	**86.2 / 92.3**	**87.7 / 94.1**	**1.42**
Area 5 (38 buildings; 530 m x 600 m)				
TUM	**85.0 / 81.1**	**81.6 / 88.2**	**88.6 / 90.9**	**1.55**

Table 1. Evaluation of the building detection results. The column headings are explained in Section 2.

Area 3: This area shows a similar distribution of the area-based quality metrics for buildings as area 1. The object-based metrics show a correctness of 100% for all methods except HAN and WHU. Most of the missed buildings larger than 50 m^2 have very complex roof shapes. Except for TON, WHU, and CAL, all methods achieve a correctness Cm_{50} higher than 95% for buildings. The results for trees show similar trends as in area 2, except that the per-tree completeness and correctness values for all trees (Cm_{ob} / Cr_{ob}) are generally lower. This is due to the fact that there are more small trees and low vegetation. HAN is particularly weak in detecting small trees. All methods detect more than 88% of the trees larger than 50 m^2; TUM is the only method achieving 100% for both Cm_{50} and Cr_{50}.

Toronto: For this test area, results were only submitted for building detection by TUM. Comparing these results to the ones achieved by TUM in Vaihingen, there are much more false positives. They mainly occur at building outlines where high-rise buildings occlude other objects or cause shadows.

Discussion: A comparison of the results for the Vaihingen test sites shows that area 1 offers the least favourable conditions for automated object extraction. A combination of trees close to multi-storey buildings casting shadows on them, terrain discontinuities, and complex roof shapes including small and low appendices at different height levels causes problems for all approaches. Buildings can nevertheless be detected reasonably well, but tree detection breaks down for most methods. The most favourable conditions for automation are found in area 2, characterised by few high-rise buildings. The problem of trees in the shadow cast by buildings is alleviated by the fact that this occurs on nearly horizontal terrain. In area 3, trees close to buildings do not affect the results because they are at the same height level. In this area, the main problems are complex roof shapes and a large amount of low vegetation, the latter causing problems for tree detection. The difference in the performance of TUM for Vaihingen and Toronto indicates the limitations of a traditional stereo configuration in the presence of high-rise buildings causing large occluded areas.

Name	Cm_{ar} / Cr_{ar} [%]	Cm_{ob} / Cr_{ob} [%]	Cm_{50} / Cr_{50} [%]	*RMS* [m]
Area 1 (105 trees)				
CAL	37.2 / **80.1**	30.5 / 53.9	60.0 / **100.0**	1.51
HAN	41.4 / 69.2	27.6 / 46.1	44.4 / 40.0	1.58
LJU	59.3 / 61.8	**63.8** / 47.2	80.0 / 90.9	1.65
TUM	**69.3** / 71.2	61.0 / **58.3**	**90.0** / 90.9	**1.12**
WHU	43.9 / 63.1	43.8 / 46.5	50.0 / 90.0	1.15
Area 2 (162 trees)				
CAL	**91.4** / 60.7	**91.4** / 45.8	**100.0** / 85.0	**1.12**
HAN	74.0 / 73.1	58.0 / **86.6**	91.7 / 93.3	1.30
LJU	88.9 / 59.2	79.0 / 55.2	**100.0** / 77.0	1.18
TUM	72.0 / **78.5**	63.0 / 82.4	**100.0** / **100.0**	1.47
WHU	64.2 / 71.5	48.8 / 70.9	91.7 / 94.7	1.34
Area 3 (155 trees)				
CAL	**83.8** / 58.6	**81.3** / 28.1	**100.0** / 87.0	1.23
HAN	55.9 / 77.0	29.0 / 68.9	89.5 / **100.0**	**1.10**
LJU	76.7 / 58.7	70.3 / 39.1	**100.0** / 81.0	1.25
TUM	69.5 / **80.1**	53.5 / **76.4**	**100.0** / **100.0**	**1.10**
WHU	50.3 / 67.6	32.9 / 55.1	88.9 / 76.5	**1.10**

Table 2. Evaluation of the tree detection results. The column headings are explained in Section 2.

The RMS errors for buildings are in the range of 1-2 times the average point distance of the ALS data (used by all methods except FIE). Of the methods using special techniques for approximating the building outlines (FIE, LJU, TON, VSK), only LJU is consistently among the best performers, but some approaches just relying on morphological operators for smoothing the classification results are in a similar range. The full accuracy potential given by the image data has obviously not been exploited by any of the compared methods. The RMS errors for trees are in the order of magnitude of the reference.

Comparing the methods based on their processing strategies, no general trend can be observed. The data may indicate that one drawback of the supervised methods used in the test is their lack of incorporating a global view of the data. The best model-based approaches for building detection apply segmentation, thus using larger entities as the basis for classification. TUM (supervised) has problems at building boundaries because each pixel is classified independently. HAN models local context by CRF, but no long-range interactions are considered. On the other hand, VSK shows the problems of a segment-based approach when segmentation fails due to small roof structures.

Of the methods only using ALS points (HAN, VSK, TON), HAN achieves the lowest per-object correctness for buildings and a poor performance in detecting trees, partly due to a confusion of buildings with trees having smooth canopies. This may be interpreted as an indicator for the importance of the radiometric data for vegetation extraction. VSK achieves a slightly lower completeness in building detection than other methods. This may be caused by a failure to find small roof planes. TON performs well in area 2, but has problems in more complex environments. The only building detection method entirely based on images (FIE) achieves a slightly worse area-

based correctness and RMS value than most of the other approaches, though its performance on a per-building basis is quite good. All the other approaches (CAL, LJU, TUM, WHU) combine ALS and image data. Of these methods, LJU performs consistently well in building detection in all areas, achieving the best area-based correctness and good RMS values in all cases. For trees, it produces a few more false positives in areas 2 and 3 than TUM. The building detection results for TUM are similar to LJU on a per-object level, but a bit worse in the per-area metrics. TUM is consistently the best performing algorithm for tree detection if both completeness and correctness are taken into account. CAL is in a similar range as TUM for buildings and slightly better than LJU for large trees. WHU performs worse in building detection than most of the other approaches, but achieves a similar quality in detecting trees.

In the context of road extraction, Mayer et al. (2006) state that a completeness of 70% and a correctness of 85% are required for real practical importance. By these standards, all the compared methods are practically relevant for extracting buildings larger than 50 m^2. All methods except WHU, TON, and HAN are also relevant if smaller buildings are considered. Only LJU and TUM consistently achieve these standards for trees larger than 50 m^2. All methods fail if smaller trees are also considered.

4.2 Task 2: 3D Building Reconstruction

The evaluation of the building reconstruction results is summarized in Table 3.

Area 1: The correctness of the roof planes is better than 94% for all methods, but there are large variations in completeness. ITCE misses more than 34% of the planes in both variants. The best results are achieved for CKU and YOR, who detect more than 85% of the planes. Undersegmentation is the dominant type of error, occurring in 36-42 cases ($N_{N:1}$). The quality metrics hardly change if only roof planes larger than 10 m^2 are considered. This indicates that the presence of small building structures does not simply result in more generalized building models, but may prevent the detection of the dominant planes. The best *RMS* value (66 cm) is achieved by CKU, which relies on manual measurement. The other *RMS* values (75-94 cm) have 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 (all except CKU).

Area 2: In this area, there is a clear difference between the quality metrics for all planes and for roof planes larger than 10 m^2. It would seem that except for ITCE1, most roof planes can be detected and nearly all of them are correct. There are fewer instances of undersegmentation ($N_{N:1}$ values between 3 and 7). However, for both variants of ITCE, this includes all high-rise buildings, which are reconstructed by single planes. This is reflected by the very poor *RMSZ* values achieved by that method. The other methods produce quite good reconstructions of the main roof structures; ITCX produces a few larger false positive planes with one of the few smaller residential buildings in this area. In general, both the planimetric and the height accuracy are slightly worse than in area 1.

Area 3: This area shows a similar distribution of completeness and correctness as area 1. There is also a rather small difference between the values for all roof planes and those for roof planes larger than 10 m^2. Undersegmentation occurs more frequently than in area 1, which may be explained by a large number of small attachments to the houses merged with neighbouring roof planes. Again, YOR and CKU achieve the best results. Two groups only submitted results for this area. Of these groups, BNU shows quite low completeness values, missing even some large dormers. FIE does a good job in detecting planes, but produces the highest number of false positives. The geometrical metrics (*RMS*, *RMSZ*) are also in the same range as in area 1.

Name	Cm_{ob} / Cr_{ob} [%]	Cm_{10} / Cr_{10} [%]	$N_{1:M}$ / $N_{N:1}$ / $N_{N:M}$	*RMS* [m]	*RMSZ* [m]
Area 1 (288 roof planes)					
CKU	86.7 / **98.9**	86.7 / **99.**3	10 / 36 / 3	**0.66**	0.70
ITCE1	60.8 / 94.6	58.5 / 94.0	16 / **26** / 17	0.91	0.55
ITCE2	65.3 / 97.3	63.3 / 97.3	**0** / 38 / 3	0.94	0.55
ITCX	76.0 / 94.5	72.9 / 95.1	2 / 40 / **2**	0.84	**0.53**
VSK	72.2 / 96.7	77.7 / 96.5	7 / 42 / 6	0.79	0.65
YOR	**88.2** / 98.5	**89.9** / 98.2	5 / 36 / 14	0.75	0.58
Area 2 (69 roof planes)					
CKU	78.3 / 93.10	90.0 / 93.7	8 / 4 / **0**	0.85	1.02
ITCE1	**79.7** / 73.7	**94.0** / 73.7	**0** / 7 / **0**	1.11	3.33
ITCE2	**79.7** / 95.0	**94.0** / **100**	**0** / 7 / **0**	1.16	3.31
ITCX	62.3 / 92.9	74.0 / 92.7	2 / 4 / **0**	0.79	**0.44**
VSK	73.9 / **100**	88.0 / **100**	3 / 5 / 1	1.03	0.88
YOR	73.9 / **100**	90.0 / 100	5 / **3** / **0**	**0.77**	1.04
Area 3 (235 roof planes)					
BNU	54.0 / 88.1	46.2 / **100**	1 / **39** / 2	0.89	0.63
CKU	81.3 / 98.4	82.2 / 98.3	4 / 48 / 2	**0.76**	0.65
FIE	82.6 / 83.1	81.4 / 91.2	7 / 44 / 5	0.99	0.62
ITCE1	67.7 / **100**	62.8 / **100**	**0** / 47 / 2	0.96	**0.29**
ITCE2	64.3 / **100**	55.9 / **100**	**0** / 46 / **0**	1.04	0.42
ITCX	70.2 / **100**	62.8 / **100**	1 / 48 / **0**	0.87	0.30
VSK	76.6 / 99.1	74.5 / 99.1	3 / 50 / **0**	0.84	0.38
YOR	**84.7** / **100**	**89.0** / **100**	2 / 51 / 1	0.77	0.35
Area 4 (967 roof planes)					
CKU	68.8 / 80.2	72.8 / 79.5	42 / **74** / 86	1.62	N/A
YOR	**75.5** / **97.5**	**83.5** / **97.5**	**27** / 109/ **19**	**1.00**	**2.88**
Area 5 (640 roof planes)					
CKU	**70.2** / 83.3	85.2 / 84.3	11 / **44** / 43	1.68	N/A
YOR	64.4 / **85.8**	**86.1 / 85.7**	**4** / 58 / **24**	**1.07**	27.22

Table 3. Evaluation of the building reconstruction results. The column headings are explained in Section 2.

Toronto: For areas 4 and 5 we only received two results. As area 3, these areas mostly consist of flat roofs, but with a larger height variation and shape complexity. For both methods, completeness and correctness are worse than in Vaihingen. Undersegmentation, but also clusters of *N:M* relations are the dominant error source for segmentation. The varying, hardly symmetric shapes are the main reason for this and also for the higher values of *RMS*. The relatively high *RMSZ* values for YOR reflect segmentation errors; the extreme value in area 5 is caused by wrong segmentation of the roofs of several high-rise buildings. For CKU, no *RMSZ* values are presented because their heights were based on wrong orientation parameters.

Discussion: Comparing the five test areas, it would again seem that area 2 offers the most favourable conditions for automatic roof reconstruction. In the other areas, complex roof structures cause a considerable amount of segmentation errors. 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 shows that methods for roof plane detection still have room for improvement, independently from the data source used. Only CKU, FIE, VSK and YOR achieve the standards required for practical relevance according to Mayer et al. (2006) in all areas. Comparing results from semi-automatic approaches (ITCEX, CKU) to the others we cannot observe any significant difference, neither in roof plane extraction nor in accuracy.

5. CONCLUSION

In this paper, several methods from current research in urban object extraction were compared based on a benchmark data set. 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 showed 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 high-quality 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 will remain available beyond the ISPRS Congress in Melbourne. Results are continuously received and evaluated. It is the goal of these efforts to provide a reference data set as a basis for making current and future developments in urban object extraction more comparable.

ACKNOWLEDGEMENTS

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 (www.siradel.com). The authors would like to acknowledge the provision of the Toronto data set by Optech Inc., First Base Solutions Inc. and York University. The reference for Toronto was created by York University.

REFERENCES

Bulatov D., Solbrig P., Gross H., Wernerus P., Repasi E., Heipke C., 2011. Context-based urban terrain reconstruction from UAV-videos for geoinformation applications. *IAPRSIS* XXXVIII-1-C22 (on CD-ROM)

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

Champion, N., Rottensteiner, F., Matikainen, L., Liang, J., Hyyppä, J., Olsen, B., 2009. A test of automatic building change detection approaches. *IAPRSIS* XXXVIII-3/W4:145-150.

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.

Grigillo, D., Kosmatin Fras, M., Petrovič, D., 2011. Automatic extraction and building change detection from digital surface model and multispectral orthophoto. Geodetski vestnik 55: 28-45.

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. 08-007r1 http://www.opengeospatial.org/standards/citygml (10/01/2012).

ISPRS, 2012. Web site of the ISPRS test project on urban classification and 3D building reconstruction. http://www.itc.nl/ISPRS_WGIII4/tests_datasets.html (10/01/2012).

Jwa, Y., Sohn, G., Cho, W. and Tao, V., 2008. An implicit geometric regularization of 3D building shape using airborne LiDAR Data. *IAPRSIS* XXXVII-B3A:69-76.

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. *IAPRSIS* XXXVI-3/W19:227–232.

Liu, C., Shi, B., Xuan, X., Nan, L., 2012. LEGION segmentation for building extraction from LiDAR data. Accepted for publication in *IAPRSIS* XLIX-B3.

Liu, X., Wang, D. L., 1999. Range image segmentation using a relaxation oscillator network. IEEE Transactions on Neural Networks 10 (3):564-573.

Mayer, H., 2008. Object extraction in photogrammetric computer vision. ISPRS J. Photogrammetry & Remote Sens. 63(2):213-222.

Mayer, H., Hinz, S., Bacher, U., Baltsavias, E., 2006. A test of automatic road extraction approaches. *IAPRSIS* XXXVI-3: 209–214.

Moussa, A., El-Sheimy, 2012. A new object based method for automated extraction of urban objects from airborne sensors data. Accepted for publication in *IAPRSIS* XLIX-B3.

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

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 J. Photogrammetry & Remote Sens. 66(2) 157-165.

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 J. Photogrammetry & Remote Sens. 66 (6) (supplement):s13-s27.

Rutzinger, M., Rottensteiner, F., Pfeifer, N., 2009. A comparison of evaluation techniques for building extraction from airborne laser scanning. IEEE J. 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 J. Computer Vision, 47(1/2/3):7-42,

Sohn, G., Huang X. and Tao, V., 2008. Using binary space partitioning tree for reconstructing 3D building models from airborne LiDAR data. PE & RS 74(11):1425-1440.

Zhang, W., Grussenmeyer, P., Yan, G., Mohamed M., 2011. Primitive-based building reconstruction by integration of Lidar data and optical imagery. *IAPRSIS* XXXVIII-5/W12.