ISPRS Journal of Photogrammetry and Remote Sensing 87 (2014) 78–92

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Fusion of airborne laserscanning point clouds and images for supervised and unsupervised scene classification

Markus Gerke [a,*], Jing Xiao [b]

[a] University of Twente, Faculty of Geo-Information Science and Earth Observation (ITC), Department of Earth Observation Science, P.O. Box 217, 7500AE Enschede, The Netherlands
[b] Wuhan University, National Engineering Center for Multimedia Software, School of Computer, Hubei 430079, PR China

ARTICLE INFO

Article history:
Received 3 July 2013
Received in revised form 29 October 2013
Accepted 30 October 2013

Keywords:
Visibility
Segmentation
Supervised
Unsupervised
Classification

ABSTRACT

Automatic urban object detection from airborne remote sensing data is essential to process and efficiently interpret the vast amount of airborne imagery and Laserscanning (ALS) data available today. This paper combines ALS data and airborne imagery to exploit both: the good geometric quality of ALS and the spectral image information to detect the four classes buildings, trees, vegetated ground and sealed ground. A new segmentation approach is introduced which also makes use of geometric and spectral data during classification entity definition. Geometric, textural, low level and mid level image features are assigned to laser points which are quantified into voxels. The segment information is transferred to the voxels and those clusters of voxels form the entity to be classified. Two classification strategies are pursued: a supervised method, using Random Trees and an unsupervised approach, embedded in a Markov Random Field framework and using graph-cuts for energy optimization. A further contribution of this paper concerns the image-based point densification for building roofs which aims to mitigate the accuracy problems related to large ALS point spacing.

Results for the ISPRS benchmark test data show that to rely on color information to separate vegetation from non-vegetation areas does mostly lead to good results, but in particular in shadow areas a confusion between classes might occur. The unsupervised classification strategy is especially sensitive in this respect. As far as the point cloud densification is concerned, we observe similar sensitivity with respect to color which makes some planes to be missed out, or false detections still remain. For planes where the densification is successful we see the expected enhancement of the outline.

1. Introduction and Related Work

Reliable detection and delineation of urban objects in airborne remote sensing data are still ongoing and relevant research topics. Automation of data interpretation is indispensable in order to process the ever increasing number of digital airborne or satellite imagery and airborne laserscanning data (ALS). In a practical production workflow, however, methods will only be accepted if they deliver complete and correct results.

In this context, the ISPRS benchmark test on urban object detection and reconstruction (Rottensteiner et al., 2012; ISPRS WG III/4, 2013) offers a unique possibility to compare state-of-the-art methods and to identify common strengths or weaknesses and subsequently stimulate the development of enhanced algorithms.

In this paper, we focus on the detection of the four major object classes in urban environments, namely *buildings, trees, vegetated ground and sealed ground*. All these classes are relevant for topographic mapping tasks, at least in a pre-processing step, like the detection of sealed ground to facilitate follow-up processes like road extraction.

For an easier analysis of existing approaches, it is convenient to distinguish between the detection task as such and the accurate estimation of object outlines.

1.1. Detection

Many methods that have been proposed for urban object detection from airborne sensor data either use point clouds from ALS or from dense image matching (multiple view stereo, MVS) solely, for some relevant work see (Dorninger and Pfeifer, 2008; Bulatov et al., 2012; Lafarge and Mallet, 2012; Niemeyer et al., 2012). Despite the fact that point clouds alone are well suited for the detection and classification of buildings and trees, one cannot reliably distinguish between sealed and vegetated areas at ground level, especially when no reflectance or full wave information is available in case of ALS.

* Corresponding author. Tel.: +31 534874522.
E-mail addresses: m.gerke@utwente.nl (M. Gerke), jing@whu.edu.cn (J. Xiao).

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.2013.10.011

Research on fusion of height and image information leads to better detection and separation of the named classes, see e.g. (Zebedin et al., 2006; Rottensteiner et al., 2007; Khoshelham et al., 2010; Awrangjeb et al., 2012; Grigillo and Kanjir, 2012; Wei et al., 2012), where the latter two are also contributing to the ISPRS benchmark test. However, looking at the details some common open issues can be identified.

- *Pixelbased vs. segment-/object-based classification*: The question of the entity which is to be classified has long been discussed in literature. There are two main strategies: (a) classify each instance of the smallest available entity, like each pixel or point and – if desired – smooth and group the finally labeled data in a post-processing step, and (b) in a pre-processing step, group all of those smallest entities to larger compounds having similar properties and hence probably belong to the same class and perform a classification on those segments, possibly followed by a grouping, e.g. according to geometric features. The latter strategy is also referred to as object oriented image analysis (Blaschke, 2010). The segment- or object-based approach gives better results in man-made scenes and if very high resolution sensor data is available (Khoshelham et al., 2010).
  To our knowledge, however, there is not yet an approach available which exploits both 3D geometric and spectral information for a proper segmentation. Barnea and Filin (2013) do this, but for terrestrial laserscanning scenes, where one scan per station is converted into a depth image and fused with a co-registered RGB image. This method, however, treats the problem basically in the 2.5D space, and therefore an extension to 3D is not possible straightforward.
- *Supervised vs. unsupervised classification*: Another comparison criterion is whether the authors develop a supervised or unsupervised, where no training data needs to be provided. Both strategies have their own advantages. While the first group of approaches are more flexible regarding data and feature quality and selection, the latter techniques can work autonomously. Khoshelham et al. (2010) compare different supervised techniques for fused ALS and image data and Rottensteiner et al. (2007) analyse an unsupervised method based on the Dempster-Shafer theory of evidence (Gordon and Shortliffe, 1990). Recently, Lafarge and Mallet (2012) applied a Markov Random Field (MRF)-based on optimization technique, using the graph-cut framework (Boykov et al., 2001) for object detection in ALS or MVS data. The latter approach combines spatial neighborhood with geometric features, and so far no literature is known to us which employs this successful strategy to fused image and height data.
- *Geometry*: If only ortho rectified images are available, the quality of height and image data fusion is hampered by relief displacement, cf. (Rottensteiner et al., 2007). If high-rise buildings are available in the scene, a proper co-registration of height and image data, i.e. using either a true ortho image or the original perspective images, is indispensable.

From this brief overview we can conclude that for a complete and correct detection of urban objects in high resolution remotely sensed data, a combined use of height and spectral information would lead to better results, compared to a sole use of either source. Although previous work showed that a pre-segmentation of the data will help in classification, the incorporation of both, height and spectral information in the segmentation has not been demonstrated so far. As far as the classification strategy – supervised vs. unsupervised – is concerned, we can see interesting developments from the machine learning and computer vision community. In some earlier work we already used those methods (Gerke and Xiao, 2013), but to the best of our knowledge a comparison of supervised and MRF-based classification method has not been done yet for fused ALS and airborne image data.

1.2. Building outline estimation

The delineation of 2D building outlines is not done explicitly in most approaches; the classified pixels or rasterized ALS points of the respective object are directly interpreted as outlines. For pixel-based detection methods this quite simple technique leads to unrealistic “zig zag” outlines (Wei et al., 2012). If a pre-segmentation of the data is performed, the shape of the final object boundary is probably better (smoother) but in particular if the segmentation is based on ALS data, the accuracy of the outline location and shape depends on the actual point spacing (Sampath and Sha, 2007).

More sophisticated techniques like the ones presented by Dorninger and Pfeifer (2008), Sampath and Shan (2010) or Brédif et al. (2013) apply some advanced regularization techniques to better represent the overall building shape, but still the final quality depends upon the point spacing in case ALS data is used. Kim and Habib (2009) use the original images to enhance the outline of building models, but they restrict themselves to pre-defined building primitives.

In (Xiao, 2012) we propose a new technique to locally densify a point cloud using images. Given that the ground sampling distance is smaller than the average point spacing of ALS point clouds, also assuming sub-pixel accuracy for low-level image operators, this technique should also obtain more accurate building outlines for the fusion of airborne images and ALS data, compared to ALS-only techniques.

1.3. Scope and contribution of this paper

In this paper we present a new method which consequently integrates ALS and large frame camera images to exploit on the one hand the very accurate, homogeneous and complete 3D geometry from the point cloud, and on the other hand the spectral information from the images to detect urban buildings, trees, natural and sealed ground objects.

Two improvements compared to other works are developed. The first one concerns the introduction of a new advanced segmentation strategy which already exploits urban object properties. The second one is an extension of the HPR (Hidden Point Removal) operator (Katz et al., 2007) for a visibility check needed to combine color information from airborne images and from the ALS point clouds.

We use the same features to compare and evaluate a supervised approach, based on Random Trees (RTrees) classification (Breiman, 2001), and a fully automatic method, based on graph-cut optimization (Boykov et al., 2001).

Furthermore, we describe an enhancement of our method to plane-based point cloud densification using the full image information, which was initially introduced for oblique views (Xiao, 2012).

2. Data Preparation

We assume irregular ALS data and images including proper orientation and calibration information are available. In order to prepare the data for further processing, we (1) filter the ground points and compute normalized heights for non-ground points, (2) apply spatial enumeration, that is we convert the point cloud into a voxel representation, and (3) determine which points are visible per image.

2.1. Ground filtering and height normalization

One obvious property of buildings and trees as compared to other objects in urban scenes is that they are significantly elevated above the ground, thus the so-called normalized height – height above ground – is a feature which is used in most urban classification approaches. We use the tool *lasground*, being part of *lastools* (Rapidlasso, 2013), to label each point whether it is a ground point or not. The software largely implements the method proposed by Axelsson (2000), i.e. it is based on mesh simplification. In a subsequent step for each off-ground point the height difference between that point and the closest ground point is computed and stored as the normalized height.

2.2. Spatial enumeration

Another pre-processing step is to perform a voxelization of the point cloud. The motivation for this is to achieve a more regular point pattern for ALS data. This is of particular importance in strip overlapping areas. However, the point cloud segmentation and the point-based geometric features such as normalized height and plane normal related measures as introduced below are computed from the original data. In this sense the voxels only carry the information derived from the points inside a particular cube. The voxel cube side length is defined in order to ensure a good sampling of the original data, i.e. it is adjusted to the approximate point spacing.

2.3. Visibility analysis

For the fusion of image information with the 3D point cloud, we need to perform a visibility analysis to ensure that each ALS point retrieves the spectral information from the correct pixel in an individual airborne image. In the literature we basically find three different categories of approaches to solve that problem: surface-based, voxel-based and based on a convex hull. The first type of methods, where one applies ray tracing methods using surfaces to detect obstruction, is reliable, as long as one can derive a correct (3D) mesh of the point cloud. However, especially in the case of ALS data where for instance building walls are only sparsely or not represented at all, this is not possible without having additional semantic information available. The same holds for voxel-based techniques (Nyaruhuma et al., 2012), where obstruction is identified relying on a voxel representation. For both types of approaches points in the background might be labeled as being visible because the input data only insufficiently represents the scene. The Z-buffer method, which is used in computer graphics for scene rendering would have the same problem, since it also relies on a closed surface object representation (Amhar et al., 1998). An interesting approach which does not need any expensive pre-computations and works with point cloud data directly is presented by Katz et al. (2007): the so-called HPR (Hidden Point Removal) operator. The idea behind this method is that if we reflect the point cloud with respect to a sphere which is centered at the observer (image projection center in this case), all points located on the convex hull of the flipped point cloud are assumed visible. This algorithm also works reliably for ALS, however, the critical step is the approximation of R – the radius of the sphere. If it is too small concave structures might not be projected on the convex hull, and if it gets too large background points might be placed on the hull. In order to find a good R we implement a similar strategy as proposed by Katz et al. (2007): a second virtual camera is placed opposite of the actual one with respect to the center of gravity of point clouds, and we find an optimal R by maximizing the difference of number of points visible from both cameras: imagine a simple horizontal plane. In that case all points will be visible, regardless of whether the observer is above or below, so the difference of visible points will be zero. However, if there are some structures on top of the plane, only the camera placed above will "see" (may be only parts of) it, while for the camera from below only the ground points remain visible, so the difference of number of visible points is at a maximum. To reduce the computation time, this visibility analysis is done once per image, and the visible points per image are stored individually and used later on.

See Fig. 1 for an example of the implemented visibility check. Part (a) shows the complete ALS point cloud while (b) shows a cut-out of one aerial image. The encircled regions depict sample areas from the ALS which are occluded by the tall buildings in the image. In part (c) the filtered point cloud is shown, where only ALS points are depicted which are visible in the image. The point color is grabbed from the image (nearest neighbor).

3. Method

The workflow of processes within the method we propose is sketched in Fig. 2. The input is given by the original ALS data, its voxel representation (see Section 2.2) and the airborne images. Features are computed from the point cloud and from the images and assigned to each voxel. After the segmentation, which uses color and geometry, the feature values are assigned to the respective segments. We then propose two different independent classification schemes: a supervised approach, based on RTrees, and an unsupervised approach which applies a graph-cut-based

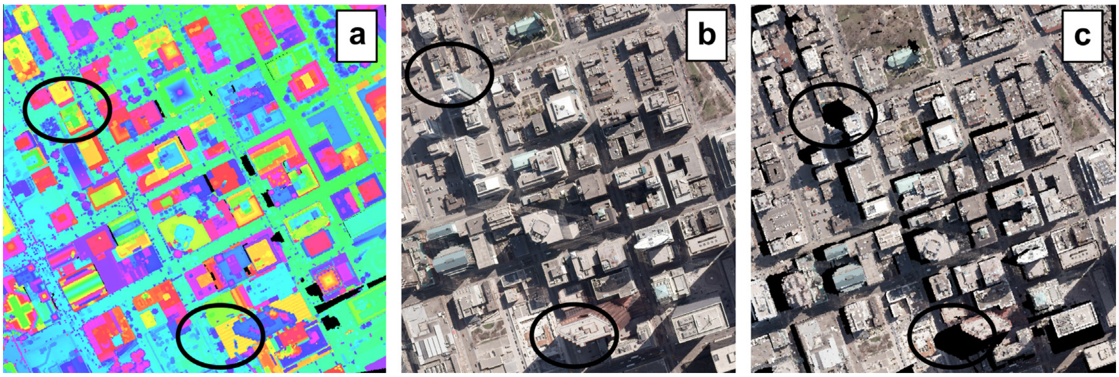


Fig. 1. Example for visibility check, images from the Toronto dataset, see Section 4.2 for a description. (a) Complete ALS point cloud, (b) aerial image (cut out), (c) only points corresponding to visible parts in the image, color grabbed from the image, black parts (also circles): non-visible areas. For the colored figure please refer to the online version of this article.

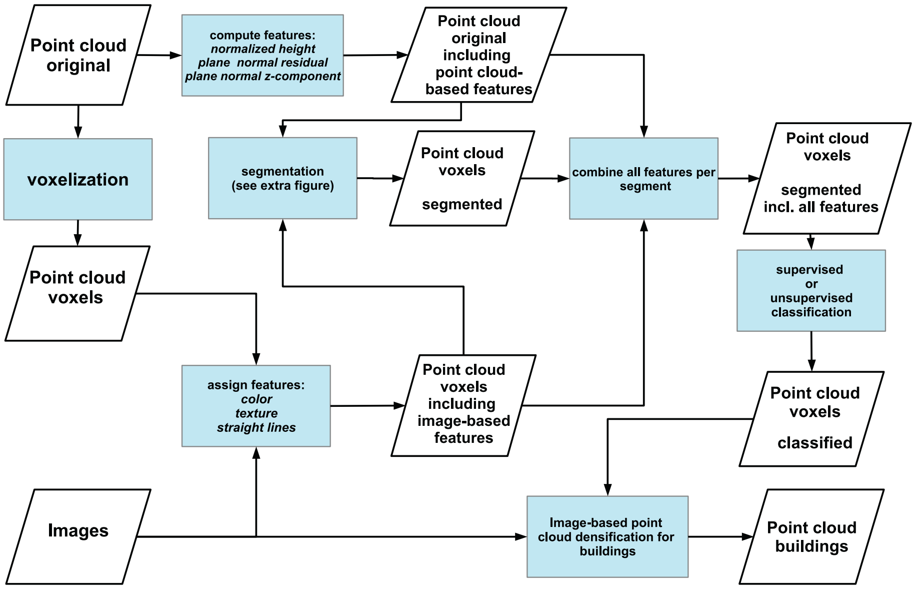

Point cloud original
compute features: normalized height plane normal residual plane normal z-component
Point cloud original including point cloud-based features
voxelization
segmentation (see extra figure)
Point cloud voxels segmented
combine all features per segment
Point cloud voxels segmented incl. all features
Point cloud voxels
supervised or unsupervised classification
assign features: color texture straight lines
Point cloud voxels including image-based features
Point cloud voxels classified
Images
Image-based point cloud densification for buildings
Point cloud buildings

Fig. 2. Entire workflow from point clouds and images to classification and building outlining.

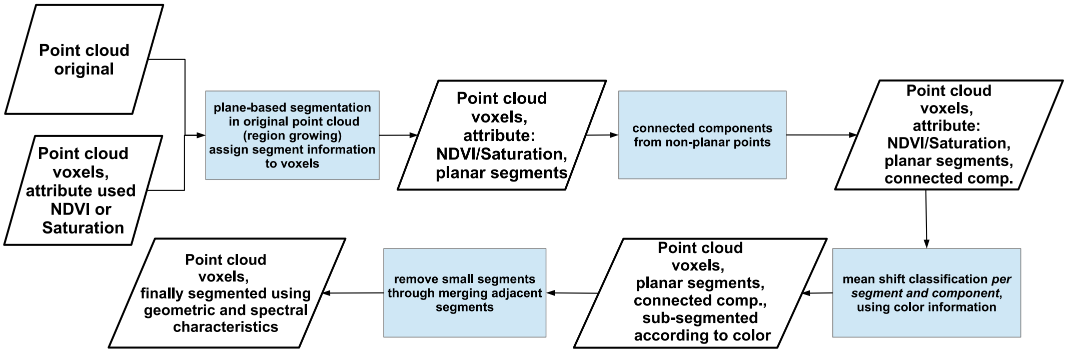

Point cloud original
Point cloud voxels, attribute used NDVI or Saturation
plane-based segmentation in original point cloud (region growing) assign segment information to voxels
Point cloud voxels, attribute: NDVI/Saturation, planar segments
connected components from non-planar points
Point cloud voxels, attribute: NDVI/Saturation, planar segments, connected comp.
mean shift classification per segment and component, using color information
Point cloud voxels, planar segments, connected comp., sub-segmented according to color
remove small segments through merging adjacent segments
Point cloud voxels, finally segmented using geometric and spectral characteristics

Fig. 3. Flowchart: segmentation of the point cloud considering geometry and color information. For the colored figure please refer to the online version of this article.

optimization scheme. Finally, the voxel clusters labeled as buildings are fed into the method for the image-based point cloud densification.

3.1. ALS- and image-based features

One feature extracted from the ALS point cloud has already been motivated in the context of data preparation, c.f. Section 2.1: the normalized height of a point, i.e. its elevation above the closest ground point. Besides this height value, which helps to separate building roofs and trees from ground, we look at planar faces. Man-made objects are mainly composed out of planar faces, as opposed to natural surfaces like trees and shrubs. Research on building model generation from ALS has shown that those planes can be extracted accurately and – depending on the point spacing and the plane size – reliably from point clouds (Oude Elberink and Vosselman, 2011). So, we estimate per point the normal of a face, composed out of the 10 closest points in the vicinity. The number of points to be considered in the neighborhood is dependent on the point cloud density and this actual value was found to be good for the data used later. Some empirical tests revealed, however, that the overall classification result is not influenced significantly as long as this number is within reasonable limits. One feature we use is the residual of the normal, which indicates whether the surface is smooth or rough, and this is helpful to separate natural from artificial surfaces. The residual of the normal corresponds to the smallest eigenvalue of the covariance matrix associated with the center of gravity, computed from all points under consideration; the normal vector is the corresponding eigenvector. In addition, we are interested in the Z-component of the normal vector since it helps to distinguish horizontal from slanted planes. For the classification problem at hand this feature contributes to the identification of building roofs.

From the images, we compute color values (Hue, Saturation, and NDVI if infrared is available), texture in the form of a standard deviation in a 9×9 matrix around each image pixel, and straight lines. The line growing algorithm by Burns et al. (1986) is used to extract straight lines. We then encode per image pixel belonging to a straight line, the length and direction of that line as two features. Again, the incorporation of such an information into the classification will help to distinguish man-made from natural objects, since at man-made structures, such as roofs or road surfaces we find linear elements. The color and texture features help to identify vegetation.

Thanks to the visibility analysis (Section 2.3), we ensure that image based features are assigned to the correct voxel. Since we have overlapping images, values for a certain feature will be observed in multiple images. Therefore, the final feature value will be computed from the median of all input values.

3.2. Segmentation of point cloud

From the related work, we learnt that a meaningful pre-segmentation according to some basic geometric – and optimally – spectral information producing ideally an oversegmentation of the actual objects leads to better final classification results. In particular we are aiming at detecting buildings, trees, natural and sealed ground. While the majority of buildings or sealed ground areas are physically composed out of planar faces, tree (crowns) are normally not this regular. Concerning natural ground the geometric shape heavily depends on the low vegetation type (lawn, shrub). A plane-based segmentation of the point cloud would already help to separate most classes (Xu et al., 2012), however there will remain a large risk of undersegmentation. Consider two example cases: a road is connected directly to a lawn area. A plane-based segmentation would define the entire area as one entity and thus result in a wrong subsequent classification. Second, a tree is standing close to some road furniture, like lamp-posts or traffic lights and all these objects from a cluster of non-planar points. Again, feeding that as one segment into a classification would result in a wrong and inaccurate labeling.

To reach a better pre-segmentation our strategy makes besides geometric properties also use of color information, c.f. Fig. 3. In a first step the original point cloud is segmented using the region growing algorithm, proposed in (Vosselman et al., 2004), yielding a segmentation according to planarity of segments. The segmentation information is then assigned to the voxel representation. All voxels not assigned to a planar segment are clustered according to spatial distance (connected components). To further sub-segment the planar segments and connected components depending on the surface type a mean shift segmentation (Comaniciu and Meer, 2002) is applied, where the NDVI – or if NIR is not available – the saturation is used as feature.[1] Each class from mean shift defined in the feature space is assigned to the point cloud and a connected component clustering is done per class repeatedly to avoid spreading of a class over the entire area. In a final merging step small segments below a pre-defined threshold are fused with the adjacent segment which has the most similar NDVI (or saturation, respectively) value. In Fig. 4 an example is provided which illustrates the method. The original ALS point cloud is shown in (a) and (b), where the color codes the height in (a) and (b) shows the initial plane based segmentation result. The encircled area shows small garden areas in front of the building (see false color image in c) which are on the same plane as the adjacent road and thus assigned to one joint segment. After the mean shift segmentation the large plane is subsegmented (d). Although now we can observe an oversegmentation, the final classification result (e) shows a good separation of the sealed ground (turquoise) and the vegetated ground (light green). Note that for a better visualization the ALS points are drawn with a quite big diameter, leaving them dilated.

3.3. Combination of features per segment

We need to compute per feature one value per segment, because the latter one is the entity which will be classified. Therefore, we compute a mean value per feature, associated to each segment. In addition, we compute a standard deviation which is used as weight in the optimization-based classification. To summarize, the following features are available per segment:

- normalized height: helps to distinguish ground from non-ground segments,
- z-component of plane normal: helps to distinguish horizontal from slanted planes,
- residual of plane normal: a measure for segment roughness,
- 2 or 3 color features: hue, saturation, NDVI (if IR available),
- standard deviation in a 9x9 window (image): texture measure, related to surface properties,
- straight line length/direction (2 features): evidence for man-made structures.

3.4. Supervised classification of segments using Random Trees

In earlier work we compared already adaptive boosting “AdaBoost” (Freund and Schapire, 1996) and Random Trees (RTrees) (Breiman, 2001) for the supervised scene classification and found out that both approaches perform similar (Gerke, 2011). Therefore, in (Gerke and Xiao, 2013) and also in this paper we only apply the latter method. The focus of this paper is to compare supervised and unsupervised methods using the same features and thus we regard it reasonable to choose one of these methods.

Reference labels as created by an operator by annotating the original images are transferred to the point cloud using a simple backprojection of the 3D points to the images, but only considering points in the respective image which are actually visible. The feature vector per segment is then fed into a RTree learning scheme. In order to monitor the quality of learning, that is to detect overfitting, the training and prediction is done several times, where each time a different subset (about 20%) of reference data is used for the training. Since in later experiments no significant difference showed up between the single runs, only the first result is used in later sections to simplify the analysis.

3.5. Unsupervised classification in a MRF framework

The advantage of using a Markov Random Field formulation for a classification task is that it combines observations (data term) with a neighborhood smoothness constraint which helps to exploit context information (Li, 2009; Ardila Lopez et al., 2011). Among a selection of optimization methods to minimize the overall energy, the graph-cut (Boykov et al., 2001) approach showed good performance in the past, e.g. for oblique airborne image classification (Gerke and Xiao, 2013). According to the overall strategy each voxel will retrieve an individual label, and the neighborhood is defined in the 3D lattice, as well. To represent the segmentation, the features as computed per segment will be used for the respective voxels inside.

We distinguish the four main classes *building, tree, vegetated ground and sealed ground* and also add a class *background* to represent “empty” voxels.

The total E energy is composed out of the data term and a pairwise interaction term:

$$E = \sum_{p \in P} D_p(f_p) + \sum_{(p,q) \in N} V_{pq}(f_p, f_q), \qquad (1)$$

where $D_p(f_p)$ is the data energy at point (i.e. voxel) p for class f_p. V_{pq} is the pairwise interaction potential, considering the neighborhood N. In particular we use the Potts interaction potential as proposed by Boykov et al. (2001) which adds a simple label smoothness term:

$$V_{pq}(f_p, f_q) = \lambda_{pq} T, \quad \text{with} \qquad (2)$$

$$T = \begin{cases} 0, & \text{if } f_p = f_q \quad \text{and} \\ 1, & \text{else} \end{cases} \qquad (3)$$

See below for a note on the value chosen for the smoothness penalty constant λ_{pq}.

[1] We conducted several experiments, also with the *Excess of Green*-Index (Gée et al., 2008), but in the end the saturation turned out to best suit our needs here. Note that at this stage it is important to only roughly discriminate different surface types, and not to perform the actual classification.

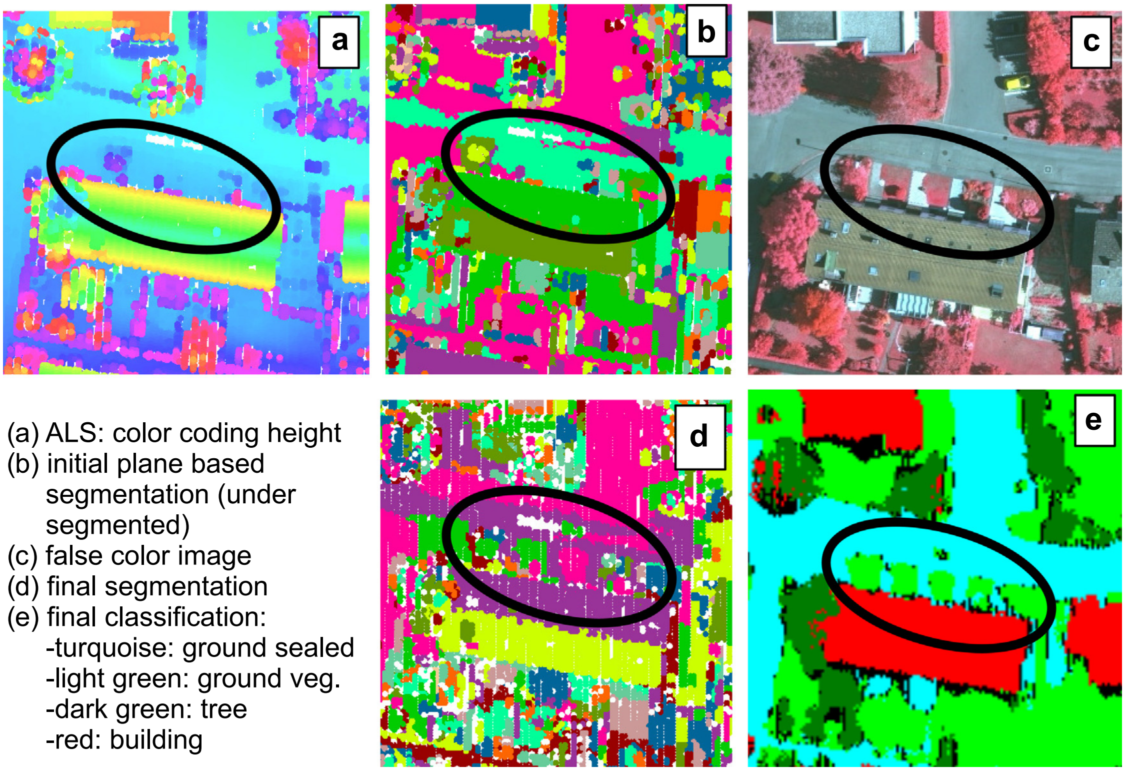

a
b
c
(a) ALS: color coding height
(b) initial plane based segmentation (under segmented)
(c) false color image
(d) final segmentation
(e) final classification:
-turquoise: ground sealed
-light green: ground veg.
-dark green: tree
-red: building
d
e

Fig. 4. Example for the proposed advanced segmentation procedure. Co-planar objects of different type get accurately separated. Images from Vaihingen dataset, see Section 4.1 for a description. For the colored figure please refer to the online version of this article.

Features contributing to the data term are normalized to the range $[0,1]$. Internally all features are stored in 8bit images, i.e. in a resolution of $1/256$. Especially for height values this is the reducing factor, and needs to be taken into account.

The feature values contribute to factors for the total energy, depending on the actual class. Each factor $S_\square$ is initialised with 1 to avoid that in case a feature does not contribute evidence for a particular class the total energy vanishes.

Normalized Height m_H (represented in dm): In the energy computation we consider that ground objects have a low height, and – depending on the object type – buildings or trees have a significantly large height above terrain. The energy is proportional to the height difference to some pre-defined threshold values, depending on the object type. For instance, for ground objects the difference to 1 m, i.e. 10 dm, is used as energy value and computed as $m_{H10} = |m_H - 10|$. The fact that we can store only 256 different values means that we can represent normalized heights up to 25.5 m. All heights above this value are set to that maximum. For our task this does not restrict the functionality since the normalized height is basically used to differentiate ground from non-ground features only.

$$S_H = \begin{cases} \min(m_{H30}, m_{H60}, m_{H90}), & \text{if } f_p = \text{building} \\ m_{H30}, & \text{if } f_p = \text{tree} \\ m_{H10}, & \text{else} \end{cases}$$

Since the height is defined in *dm*, Untitled Document 1the energy formulation for buildings is in this example best for buildings up to 9 m, but can easily extended for taller buildings.

Line Length m_{L-} and m_{L+}: If one or more lines are assigned to the segment where the voxel is located, we compute two different values: m_{L-} the difference in length to the shortest line in the overall area and m_{L+}: the difference in length to the longest line. Those two values are used in the energy computation, depending on the class assuming that longer lines can be found at buildings.

$$S_L = \begin{cases} m_{L+}, & \text{if } f_p = \text{building} \\ m_{L-}, & \text{else} \end{cases}$$

Note that if less than two lines are available within a voxel, S_L remains the initial value $S_L = 1$, i.e. it is neglected within the energy computation.

Plane Normal, Z-component m_Z: This feature is to distinguish horizontal from non-horizontal planes and the basic idea is that for building roofs we can have horizontal planes or inclined planes, while the former one is supported through a large m_Z (close to 1, thus minimum energy at $1 - m_Z$) and the latter one is expected to show m_Z values around 0.5. For sealed ground we assume horizontal planes.

$$S_Z = \begin{cases} \min(1 - m_Z, |0.5 - m_Z|), & \text{if } f_p = \text{building} \\ 1 - m_Z, & \text{if } f_p = \text{sealed\_grnd} \\ \min(1 - m_Z, |0.5 - m_Z|) + C, & \text{else} \end{cases}$$

For *tree* and *vegetated ground* areas the normal vector cannot contribute any evidence – it is arbitrary. In order to avoid the impact of this ignorance on the total energy, the factor for those classes is the same as the minimum energy contributing to the other two classes, with an added very small constant energy *C*.

Texture: Standard Deviation m_T: This texture measure is useful to characterize surface roughness. It is computed as the standard deviation in a sliding 9x9 window in an image. Since we assume a large value for trees we also compute the overall maximum standard deviation m_{Tmax}. The smaller the difference to the maximum value, the smaller also the energy for the tree-class.

$$S_T = \begin{cases} |m_T - m_{Tmax}|, & \text{if } f_p = \text{tree} \\ m_T, & \text{else} \end{cases}$$

Color m_C: If an infrared color channel is available in the image data we compute the NDVI, otherwise we use the RGB values to compute the saturation (SAT) and HUE values.

NDVI: this index is defined in $[-1,1]$, but here normalized to $[0,1]$. The closer it is to 1, the more likely it is vegetation, i.e. the respective energy should be small.

$$S_{C_N} = \begin{cases} 1 - m_{C_N}, & \text{if } f_p \in \{\text{tree}, \text{veg\_grnd}\} \\ m_{C_N}, & \text{else} \end{cases}$$

SAT and HUE: We observed that the saturation is generally high for vegetation, while it is low for sealed areas, therefore the saturation is used in a similar manner as the NDVI to compute the energy S_{C_S}. In the HUE definition the color green is defined at 120°. Hence we assume that vegetation has a peak around that hue value, while sealed areas show a relatively small signal there:

$$S_{C_H} = \begin{cases} |m_{C_S} - 120|, & \text{if } f_p \in \{\text{tree}, \text{veg\_grnd}\} \\ \min(|m_{C_S}|, |m_{C_S} - 240|), & \text{else} \end{cases}$$

Note: For the sake of simplicity the fixed angular values are given in the original unit in this equation. In practice they are also scaled to $[0,1]$.

To consider the uncertainty inherent in the feature values, we add a penalizing energy $S_{\square_{pen}}$ to *each factor* which is proportional to the standard deviation computed during the merge of feature values per segment.

So finally each feature contributes to a final factor $S'_{\square} = S_{\square} + S_{\square_{pen}}$ per object class which is defined in $[0,1]$. The energy computed from all features voting for a certain label p is

$$D_p(f_p) = S'_H(f_p) \cdot S'_L(f_p) \cdot S'_Z(f_p) \cdot S'_T(f_p) \cdot \begin{cases} S'_{C_N}(f_p), & (\text{if NDVI is available}) \\ S'_{C_S}(f_p) \cdot S'_{C_H}(f_p), & \text{else} \end{cases}$$

In the graph-cut implementation we use, the energies need to be presented as integer values. Some internal experiments showed that an actual ranking of the energy per entity gives the best result. Thus we assign an energy defined in $[0, N-1]$ – where N is the number of classes – to $D_p(f_p)$, according to the sequence in the original $D_p(f_p)$ computation. Another reason to choose this ranking is that the smoothness penalty constant λ_{pq} can be adjusted accordingly. In our experiments we set it to 1.

3.6. Image-based point cloud densification for buildings

The basic idea behind this densification method is that planar faces, initially extracted from point clouds, have a good absolute geometric accuracy, but borders are not defined well because of the restricted ALS point cloud density. At the same time image information from the very same plane can help to refine the borders by combining area based features with edge information. A basic assumption is that the GSD of images is better than the average point cloud spacing. The proposed densification is done per plane obtained from the plane-based point segmentation of the classified building points, hence it relies on the initial classification. Given that a plane on the building would mostly have homogeneous spectral representation, the source images are segmented using spectral information. Edges of roof planes normally appear straight, so that straight lines are also used to control the final region extend.

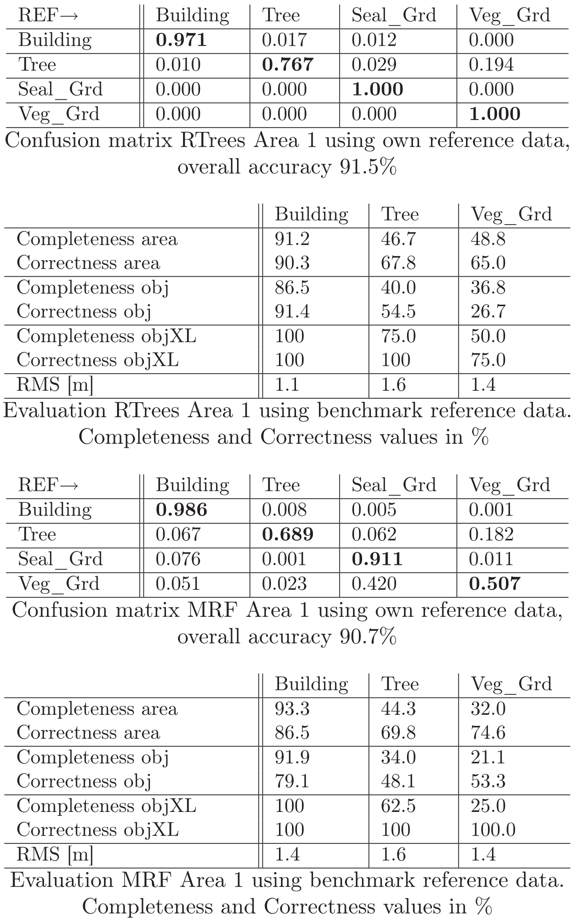
A grid space of finer resolution than the ALS data is created for the densification, 20 cm GSD in the experiments here. Each pixel in the grid space contains values from several image features. Then a graph-cut classification approach is implemented for the features combination and final classification into the two classes *building* and *background*. The energy is of the same form as defined in Eq. (1). The data energy term $D_p(f_p)$ comes from the graded value G_p, derived from image segments and the smoothness term $V_{pq}(f_p, f_q)$ represent the interaction between the roof points and the edges defined by straight lines.

$$D_p(f_p) = \begin{cases} 1 - G_p, & \text{if } f_p = \text{building} \\ G_p, & \text{if } f_p = \text{background} \end{cases}$$

We use the graph based method by Felzenszwalb and Huttenlocher (2004) for the color segmentation, which is able to preserve details in low-variability image regions but ignore details in high-variability regions. Segments in each image are selected by projecting plane points back into the image after visibility check. If the number of the points falling into the segments reached a threshold, the value of this segment C_{seg} is calculated:

$$C_{seg} = \frac{N_{pc}^2}{A_{seg}},$$

with N_{pc}: number of original ALS points falling into the particular segment, and A_{seg}: area of the image segment. Using this definition image segments which are covered by more initial points get a larger C_{seg}, i.e. it is larger for segments on the roof as for the ones covering roof and background. After calculating the values of all tested segments, C_{seg} is normalized into [0,1] by the maximum and minimum values. Then the segments from each image are resampled into the grid space. The value of each grid pixel G_p is the average of segment values from all images.



REF→	Building	Tree	Seal_Grd	Veg_Grd
Building	**0.971**	0.017	0.012	0.000
Tree	0.010	**0.767**	0.029	0.194
Seal_Grd	0.000	0.000	**1.000**	0.000
Veg_Grd	0.000	0.000	0.000	**1.000**

Confusion matrix RTrees Area 1 using own reference data, overall accuracy 91.5%

	Building	Tree	Veg_Grd
Completeness area	91.2	46.7	48.8
Correctness area	90.3	67.8	65.0
Completeness obj	86.5	40.0	36.8
Correctness obj	91.4	54.5	26.7
Completeness objXL	100	75.0	50.0
Correctness objXL	100	100	75.0
RMS [m]	1.1	1.6	1.4

Evaluation RTrees Area 1 using benchmark reference data. Completeness and Correctness values in %

REF→	Building	Tree	Seal_Grd	Veg_Grd
Building	**0.986**	0.008	0.005	0.001
Tree	0.067	**0.689**	0.062	0.182
Seal_Grd	0.076	0.001	**0.911**	0.011
Veg_Grd	0.051	0.023	0.420	**0.507**

Confusion matrix MRF Area 1 using own reference data, overall accuracy 90.7%

	Building	Tree	Veg_Grd
Completeness area	93.3	44.3	32.0
Correctness area	86.5	69.8	74.6
Completeness obj	91.9	34.0	21.1
Correctness obj	79.1	48.1	53.3
Completeness objXL	100	62.5	25.0
Correctness objXL	100	100	100.0
RMS [m]	1.4	1.6	1.4

Evaluation MRF Area 1 using benchmark reference data. Completeness and Correctness values in %

Fig. 5. Results Area 1.

The smoothness term defines the interaction between the building pixels with the edges which is represented as the straight lines in the neighborhood. Image straight lines extracted in the context of image-based feature computation are re-used. Straight lines close to selected segments are picked out. In order to project them into the grid space, they are restrained onto the processing plane. Besides the lines from images, 3D lines generated from the stereo intersection of those 2D lines are employed. Higher weights are assigned to the 3D lines since they are visible from at least three images thus treated more robust than the projected 2D lines.

Semantically, the building points should be on the same side of an edge, hence the neighborhood is defined accordingly (Lafarge and Mallet, 2012):

$$(p,q,)\in N \Longleftrightarrow \begin{cases} \text{distance}(p,q) \leqslant d \quad \text{and} \\ O(p,L) = O(q,L) \end{cases}$$

d is the maximum distance between the locations of two neighboring pixels. L is a straight line close to the pixels. $O(p,L)$ is the oriented side in which the cell p is located with respect to L.

The pairwise interaction between neighbors is formulated as:

$$\begin{aligned} V_{pq}(f_p,f_q) = {} & \epsilon \text{ if } \begin{cases} f_p = f_q \ \& \ O(p,L) = O(q,L) \\ f_p \neq f_q \ \& \ O(p,L) \neq O(q,L) \\ f_p = f_q \ \& \ L = \emptyset \end{cases} \\ = {} & 1, \text{else} \end{aligned}$$

ϵ is a pre-defined penalty value, defined in $[0,1]$. It controls the influence of lines in the classification, so a smaller ϵ gives more control to lines. Again, the graph-cut based algorithm by Boykov et al. (2001) is used to minimize the energy per plane area and ultimately by this means to find the optimal classification of grid pixels into building and background. Final building pixels are back-projected into object space using the known plane parameters.

3.7. Evaluation methods

For the evaluation of our approach we only concentrate on the ISPRS benchmark dataset (Rottensteiner et al., 2012; ISPRS WG III/4, 2013), because we believe that this is the most objective way to evaluate the performance of the method. In addition those datasets represent the diversity of European and North American City architecture quite well. Evaluation results are provided in different ways: Based on own reference data which was labeled by an operator and (partly) used to train the RTrees classifier. Those evaluation details are given in confusion matrices, which show the per-segment result. Percentages refer to the total number of reference entities, i.e. rows sum up to 100% (± because of round-off errors). The overall classification accuracy is computed as the normalized trace of the confusion matrix.

The official benchmark evaluation, which is provided here as well, does not give any indication about interclass-confusion. It gives completeness and correctness per object class in three different computations: (a) per area, i.e. independent from object on a pixel basis, (b) per object (*obj*), (c) per object, but only considering large objects (*objXL*: objects larger than 50 m^2). Results based on own reference data and given in the confusion matrix can be compared to the benchmark evaluation, but keeping in mind that own evaluations are done on a per-segment basis: a main diagonal element in the confusion matrix can be interpreted as a completeness measure since it gives the actually detected number of segments in relation to the entire number of reference segments for that particular object. The official ISPRS benchmark evaluation also provides geometric accuracy evaluation in the form of a RMSE value of distances between extracted and reference objects. Note that the ISPRS reference data does not cover the class *sealed ground*, but our own evaluation provides this information.

4. Results

4.1. Vaihingen Dataset

The testdata provided by the ISPRS benchmark test was originally produced for the DGPF camera test (Cramer, 2010). The 20 Intergraph/ZI DMC-images in the block have a 65% forward and 60% sidelap, and thus a four-fold overlap is ensured; the ground sampling distance (GSD) of the CIR images is 8 cm. A Leica ALS50 system was used for the ALS flight, the point density varies between 4 and 7 points/m^2.

The ALS and airborne image data were acquired with one month time difference (images 24 July, 2008 vs. ALS 21 August, 2008). Although minor changes in the scene might occur, we can consider the objects of interest static and unchanged.

4.1.1. Area 1: "Inner City"

The first area consists of rather old historical buildings with a complex structure. Vegetation is available, but not dominant.

See Tables in Fig. 5 for confusion matrices using own reference data and evaluation from the official ISPRS benchmark test. Both results, from RTree and MRF, are provided.

Evaluation of the supervised method using own reference data reveals a quite complete object (segment) detection result for all classes, at least 76.6% for trees, but at the same time shows an in-

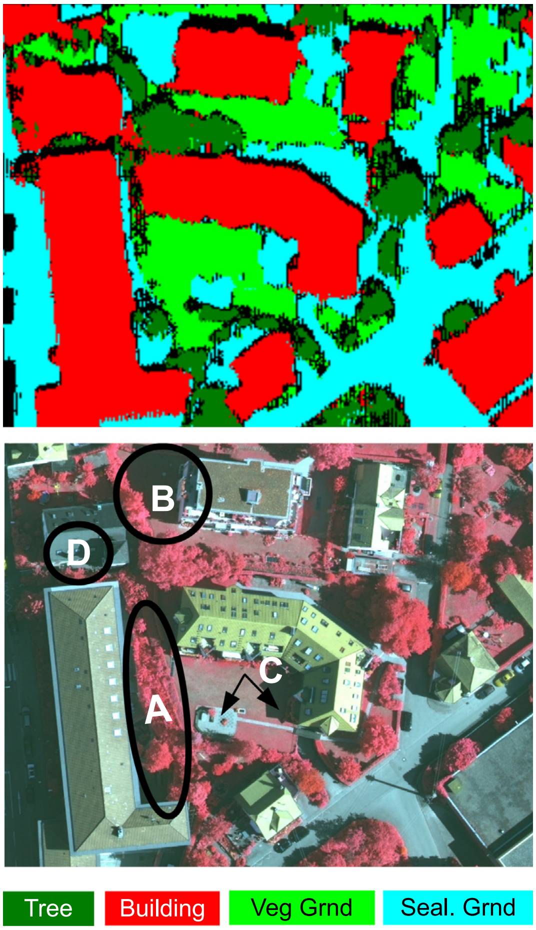


Fig. 6. Area 1: Example for the most prominent type of misclassification: trees and ground vegetation is confused (A, both RTrees and MRF), ground vegetation is classified as ground sealed (B, mainly MRF). Label image from MRF classification. Example for segmentations in (C) and (D), see text. For the colored figure please refer to the online version of this article.

ter-class confusion between trees and vegetated ground: 19% of actual trees got classified as natural ground, some 3% got classified as sealed ground. The same tendency is confirmed looking at the official benchmark evaluation. On an object basis the completeness for trees is only 40% (for large trees 75%). Refer to Fig. 6 for an example: the tree area (A) close to the building is interpreted as ground vegetation. The label image is from MRF, but in the RTrees example the similar error occurs. The reason for this is mainly in the uncertain height definition for ground vegetation: areas with shrub are considered as ground vegetation, so areas with smaller trees might then be assigned to this class.

For the unsupervised classification, we observe the same good quality for buildings, but also a similar trend concerning inter-class confusion as for RTrees. In addition the classes sealed and vegetated ground get highly confused (42% of vegetated ground got labeled sealed ground), this is also reflected in the completeness measure for vegetated ground using the ISPRS benchmark evaluation. Those confusions occur in shadow areas: in those parts the NDVI value is quite low and thus the energy in the data term in principal becomes smaller for the sealed ground class. In the supervised RTrees this problem is not this dominant since those areas are also part of the training set and thus the votes from NDVI reflectance are better adapted to the actual scene. See part (B) in Fig. 6 for a large shadow area next to the building resulting in the label *sealed ground*, although it is a vegetation area.

Notes on segmentation and outline geometry. This area can also be used to demonstrate properties of the segmentation which makes use of both, geometry and spectral properties. The large planar area (C) in Fig. 6 was initially segmented as one large segment from the region growing algorithm, but then got subdivided based on spectral properties. Because of this, the left-hand sealed area was correctly separated from the large lawn area, but – on the other hand – also the shadow area got identified as own segment. As such, this is not a problem, but such separations then increase the risk of wrong classifications and this is also visible here (same problem as in (B)). The area in (D) reveals an overall disadvantage of the presented segmentation method: since information on first and last pulse is not used here, building roof planes underneath vegetation do get split up and only the vegetation is represented in the segmentation.

The outline of segments is at the same time object boundary, given the segment is not completely surrounded by segments from the same class. The problems caused by shadow and vegetation as mentioned above also have an impact on the object boundary, which is visible in Fig. 6 as well. If, however, the building roof can well be separated from the background, see the larger buildings in the figure, the outline is well preserved. However, concerning the final outline we are affected by the ALS point density and quantification effects from voxelization.

4.1.2. Area 2: “High Riser”

The second area shows some multiple storey apartment buildings, typical for a European sub-urban area or smaller cities. Those buildings are mostly surrounded by trees and some natural ground areas.

Again, Tables in Fig. 7 show evaluation results in the same setup as before.

For this area the overall result (accuracy from own evaluation) is better as for area 1, and compared to area 1 the inter-class confusion between trees and vegetated ground is not as significant. The evaluation using the ISPRS benchmark data shows lower correctness values, especially for trees. This is because in the reference

REF→	Building	Tree	Seal_Grd	Veg_Grd
Building	**0.987**	0.009	0.000	0.004
Tree	0.009	**0.972**	0.002	0.017
Seal_Grd	0.000	0.000	**0.970**	0.030
Veg_Grd	0.000	0.000	0.000	**1.000**

Confusion matrix RTrees Area 2 using own reference data, overall accuracy 97.9%

	Building	Tree	Veg_Grd
Completeness area	94.0	83.1	69.4
Correctness area	89.0	64.3	82.2
Completeness obj	78.6	69.2	57.9
Correctness obj	42.3	76.2	62.2
Completeness objXL	100	100	83.3
Correctness objXL	100	83.7	81.3
RMS [m]	0.8	1.5	1.3

Evaluation RTrees Area 2 using benchmark reference data

REF→	Building	Tree	Seal_Grd	Veg_Grd
Building	**0.981**	0.008	0.010	0.001
Tree	0.285	**0.602**	0.058	0.056
Seal_Grd	0.005	0.003	**0.975**	0.017
Veg_Grd	0.008	0.012	0.364	**0.617**

Confusion matrix MRF Area 2 using own reference data, overall accuracy 85.0%

	Building	Tree	Veg_Grd
Completeness area	93.6	52.8	37.2
Correctness area	69.2	67.9	85.9
Completeness obj	78.6	41.7	36.8
Correctness obj	11.7	74.2	34.6
Completeness objXL	100	66.7	33.3
Correctness objXL	66.7	71.4	75.0
RMS [m]	0.9	1.5	1.3

Evaluation MRF Area 2 using benchmark reference data

Fig. 7. Results Area 2.

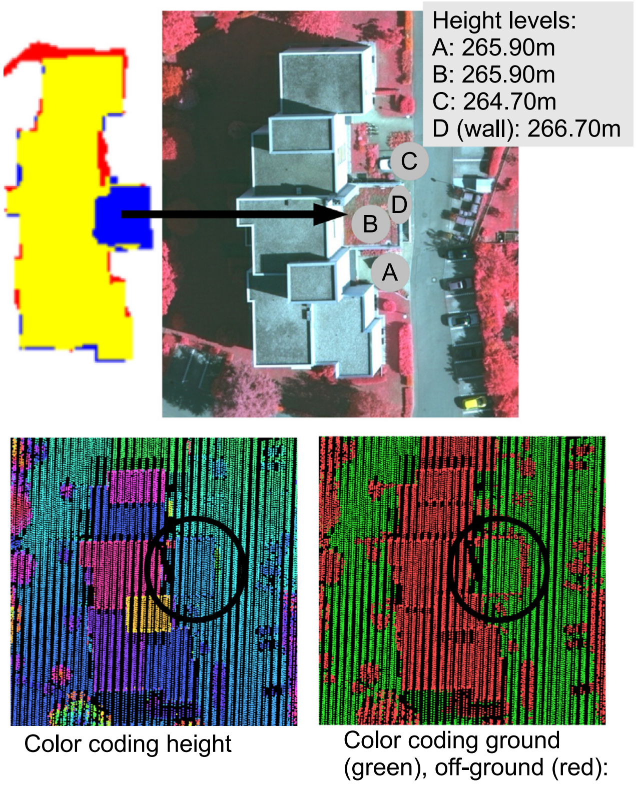


Fig. 8. Area 2: Fuzzy object definition. Upper left: ISPRS evaluation result (blue: false negative), upper right: false color image, showing the height levels around the building appendix, lower left: ALS point cloud, color coding height, lower right: color coding ground/off-ground status. For the colored figure please refer to the online version of this article.

data for the ISPRS benchmark trees are modeled as circles, while in our approach we do not fit circles, and this leads to some excess areas.

Results from the unsupervised method again show big problems related to shadow areas, and the NDVI, respectively. In this case trees and buildings largely are confused: from the own reference data it shows that almost 30% of all tree pixels get labeled as buildings, and this is also the reason for the small building correctness score when the ISPRS reference data is used. In a similar manner and for the same reason as explained for area 1, again, vegetated and sealed ground get confused.

Another issue concerns the semantic object definition. Fig. 8 shows an example building from area 2. The upper left hand image is from the evaluation, and this indicates that the appendix to this building at the eastern side is missed in the extraction (blue pixels). This appendix is a ground-level kind of backyard, surrounded by a wall. For this reason this area is assumed to be part of the building. The automatic approach, however classified it as natural ground since it is on ground level (see lower images) and covered with vegetation, see false color image. Those fuzzy object definition are clear challenges for automatic object detection.

4.1.3. Area 3: "Residential"

In this third area we mainly find single detached houses and garages with gardens and other green areas in the vicinity.

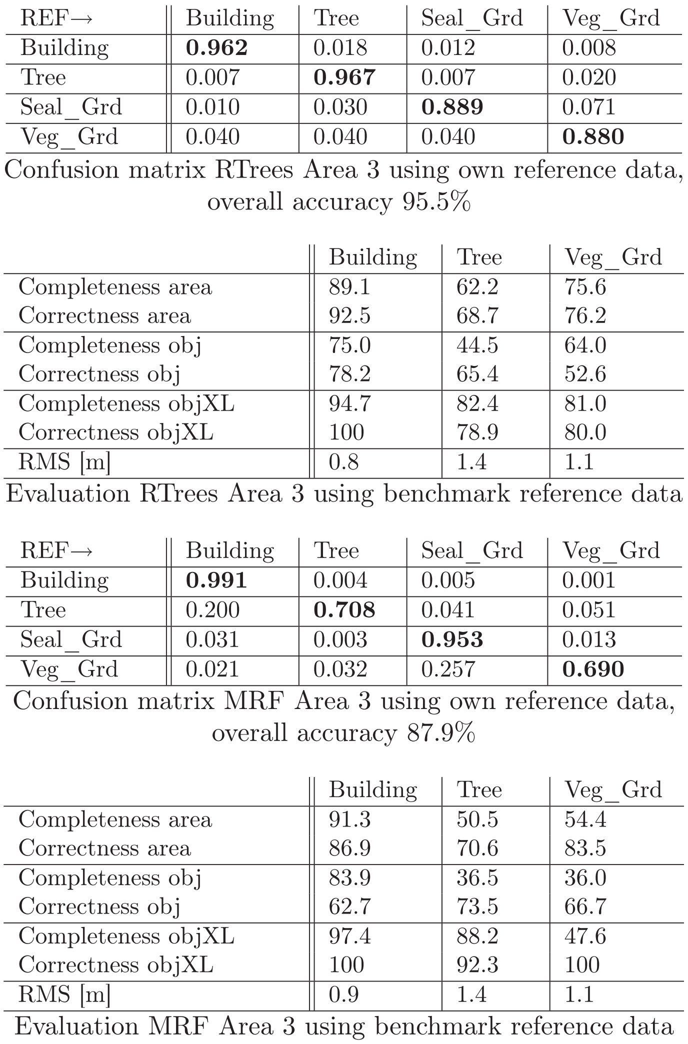
See Tables in Fig. 9 for the evaluation results.

In area 3 we can observe especially problems with small buildings/garages: some of them are missed in the classification and get labeled as ground sealed objects. In addition we note a special artifact here: there is a huge gap in the ALS point cloud, leaving large parts of one building out. Possibly this is due to the surface of the roof. Besides, in this area we can observe similar trends as in the other areas. Fig. 10 shows the gap in the point cloud (A) and also again a typical example for inter-class confusion of vegetated and sealed ground areas.

4.2. Toronto dataset

The two test areas in Toronto, Canada, are covered by 6 images, taken with a Microsoft Vexcel UltraCam-D. The forward overlap is 60%, but because of a sidelap of only 30% the two test areas are only visible in one strip entirely. The GSD is twice as large as in the Vaihingen dataset, 15cm, and only RGB spectral channels are available. The LiDAR system scanned the area using scan width of 20° which reduces the occlusion effects caused by tall buildings; the average point density is similar to Vaihingen; 6 points/m^2.

The major problem with this dataset is that between the image acquisition and the ALS flight some 2 years passed, in addition the images were captured in leave-off season while the ALS data was flown in summer. These circumstances make it very difficult to de-



REF→	Building	Tree	Seal_Grd	Veg_Grd
Building	**0.962**	0.018	0.012	0.008
Tree	0.007	**0.967**	0.007	0.020
Seal_Grd	0.010	0.030	**0.889**	0.071
Veg_Grd	0.040	0.040	0.040	**0.880**

Confusion matrix RTrees Area 3 using own reference data, overall accuracy 95.5%

	Building	Tree	Veg_Grd
Completeness area	89.1	62.2	75.6
Correctness area	92.5	68.7	76.2
Completeness obj	75.0	44.5	64.0
Correctness obj	78.2	65.4	52.6
Completeness objXL	94.7	82.4	81.0
Correctness objXL	100	78.9	80.0
RMS [m]	0.8	1.4	1.1

Evaluation RTrees Area 3 using benchmark reference data

REF→	Building	Tree	Seal_Grd	Veg_Grd
Building	**0.991**	0.004	0.005	0.001
Tree	0.200	**0.708**	0.041	0.051
Seal_Grd	0.031	0.003	**0.953**	0.013
Veg_Grd	0.021	0.032	0.257	**0.690**

Confusion matrix MRF Area 3 using own reference data, overall accuracy 87.9%

	Building	Tree	Veg_Grd
Completeness area	91.3	50.5	54.4
Correctness area	86.9	70.6	83.5
Completeness obj	83.9	36.5	36.0
Correctness obj	62.7	73.5	66.7
Completeness objXL	97.4	88.2	47.6
Correctness objXL	100	92.3	100
RMS [m]	0.9	1.4	1.1

Evaluation MRF Area 3 using benchmark reference data

Fig. 9. Results Area 3.

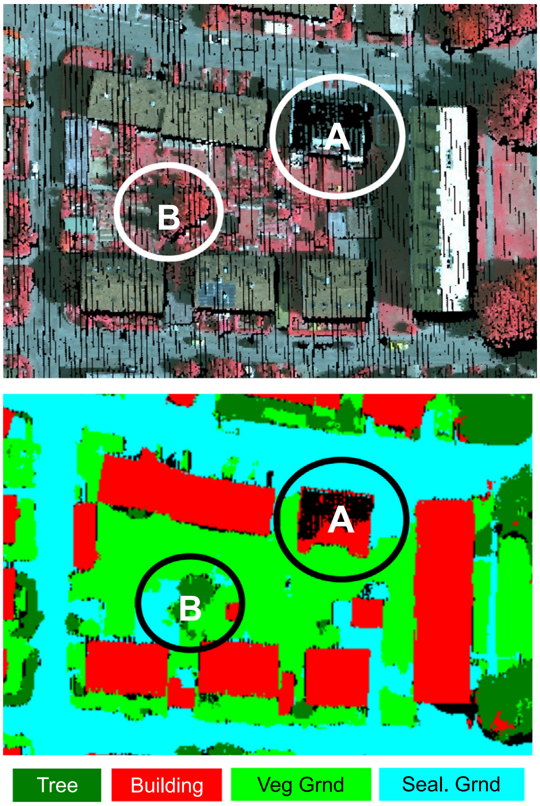


Fig. 10. Area 3: upper image: colored point cloud, lower: classification result from RTrees. Gap in point cloud (A) and example for vegetation/sealed ground confusion (B). For the colored figure please refer to the online version of this article.

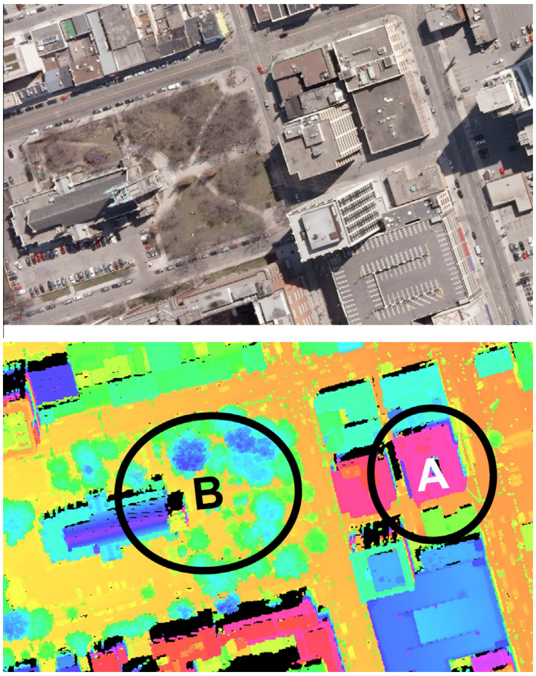


Fig. 11. Examples showing large time difference between ALS and image flight (A), also both flights were done in different vegetation periods (B). For the colored figure please refer to the online version of this article.

REF→	Building	Seal_Grd
Building	**0.894**	0.106
Seal_Grd	0.004	**0.996**

Confusion matrix RTrees Area 4 using own reference data, overall accuracy 91.3%

REF→	Building	Seal_Grd
Building	**0.927**	0.073
Seal_Grd	0.053	**0.947**

Confusion matrix MRF Area 4 using own reference data, overall accuracy 93.0%

	Building
Completeness area	71.5
Correctness area	96.8
Completeness obj	77.6
Correctness obj	57.0
Completeness objXL	78.9
Correctness objXL	92.3
RMS [m]	1.0

Evaluation RTrees Area 4 using benchmark reference data

	Building
Completeness area	80.5
Correctness area	82.1
Completeness obj	96.6
Correctness obj	22.9
Completeness objXL	96.5
Correctness objXL	67.1
RMS [m]	1.5

Evaluation MRF Area 4 using benchmark reference data

Fig. 12. Results Area 4.

tect trees or low vegetation, further some buildings changed in between the two acquisitions. Fig. 11 shows in part (A) that tree crowns are fully represented in the ALS point cloud, while in the images only the branches of deciduous trees are visible, part (B) shows a building which got extended in height significantly: in the ALS it shows a height similar to the adjacent one (see color code), while in the images it is much smaller.

Because of the mismatch between the datasets the classes *trees* and *vegetated ground* are not considered here further.

4.2.1. Area 4: Mix of high and low, complex shapes

In area 4, defined in the Toronto dataset, we find a mixture of tall and low rise buildings of different shape types; there are quite a few tall towers, as well as a church and a park area in the scene.

See Tables in Fig. 12 for the evaluation results.

The completeness for buildings in RTrees is lower than in MRF, and for correctness values we observe the opposite. This is a very interesting observation and can be explained as follows. The relatively small number of low-rise buildings causes in this case the supervised, RTrees-based, classification to misinterpret those as ground objects, while in the MRF, even if the height above ground is only some meters, those get classified correctly as buildings. This is the main reason why some buildings are missing in the RTrees result, and hence for a lower completeness. The confusion matrix from the own evaluation confirms this explanation since more building segments get labeled as sealed ground in the RTrees than in the MRF-based case.

On the other hand, especially the MRF-based approach relies on color information to distinguish vegetated from non-vegetated areas, but in this dataset there is not only a mismatch between ALS and image information due to a time difference, but also the RGB color information seems not to be sufficient to separate the classes. This is why many trees get classified as buildings in the MRF-based approach, and thus the correctness of the buildings class is lower than for the RTrees case.

In Fig. 13 some examples are given, which are used to explain some more difficulties. In the lower area of the classified label image (I, A), one can observe a varying point density which hampers detection. Parts B and C demonstrate that the human-made reference can contain errors. Actually those regions are not buildings, but yards, partly with vegetation in (C). Those areas were correctly labeled by the RTrees-based classification, but it has been assessed as being false negative pixels. In contrast, the large low-rising building with a height above ground of some 9 m, was not detected by RTrees (see above for an explanation) and hence correctly indicated as false negative.

4.2.2. Area 5: Skyscrapers

In this area we mainly find tall towers, which are partly connected by flat roof buildings. Looking at the evaluation for this area

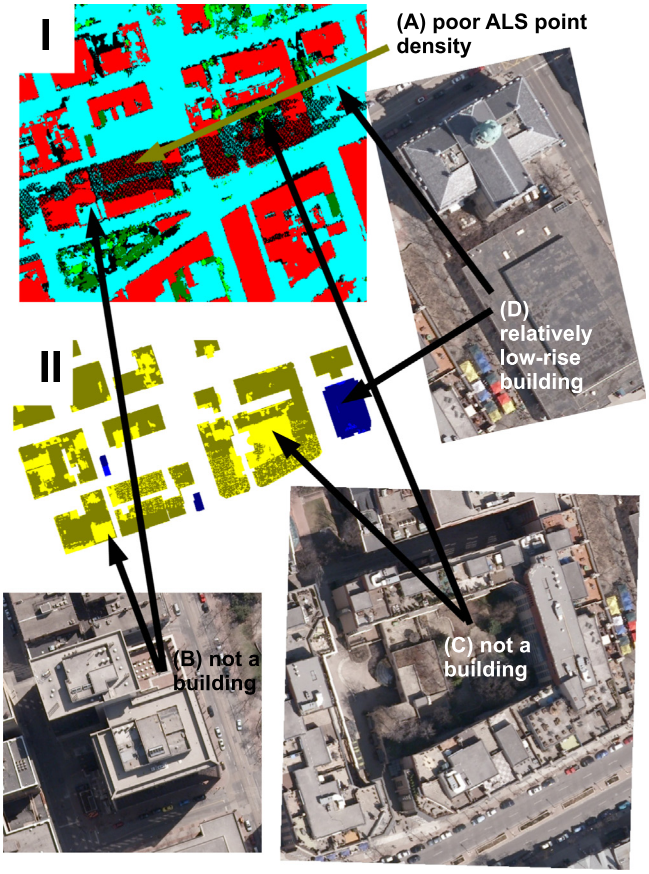


Fig. 13. Area 4: I: labels from RTrees classification (buildings and sealed ground), II: evaluation from ISPRS benchmark, yellow: false negative pixels (but object detected), green: true positive pixels, blue: false negative pixels (and object missed). (A) Shows that the point density is varying in this area, (B and C) show errors in the reference: those spots actually show no buildings, (D) low-rising building which is missed in the RTrees result. For the colored figure please refer to the online version of this article.

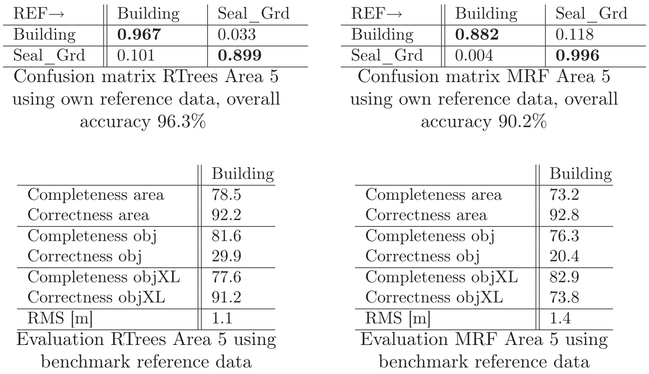
in Fig. 14 we basically observe the same trend as in area 4: better completeness for large buildings by the unsupervised approach, but better correctness from the supervised method. The missing buildings in the supervised method are all large area halls, but having only a low height. The reason for missing them is the same as above.

4.3. Image-based point cloud densification for buildings

We applied the point cloud densification to Vaihingen, areas 1 to 3, based on the supervised building classification result. In Table 1 the ISPRS-benchmark evaluation for the buildings is shown, and per item the change to the respective values from the supervised method is indicated. While the correctness increases tremendously, first of all in the object based evaluation part, the completeness decreases in most cases. Since only the previously regions labeled as building are considered, the completeness cannot increase, and because some parts get removed, the correctness does increase at the same time. In contrast to theoretic expectations the geometric accuracy measured as the RMS error of distances between reference and extraction boundary does not improve compared to the ALS-based segmentation.

The building parts indicated in Fig. 16A and D demonstrate typical problems related to the initial ALS point cloud density. At the upper building the plane is quite narrow and in (D) the overall size is small. For this reason the planes got skipped for the densification. Examples for removed false building voxel clusters are shown in Fig. 15B. The evidence from segmentation in the respective image regions is not homogeneous, and therefore the data term will have lower energies for the *background* class. For the same reason, but with a negative outcome, shadow areas, such as (B) in Fig. 16 or (A) in Fig. 15, remain false in the result. In shadow areas the segmentation is homogeneous and therefore energy for *building* are lower. The building (C) in Fig. 16 was missing already from the RTrees classification (due to its relatively low height it was labeled as sealed ground), and because only building voxels are considered here, it remains missing.

The geometric accuracy is in general worse compared to the former classification, mainly because the mentioned missing planes cause large deviations. On the other hand, if planes are completely represented they fit quite well to the reference, see for instance edges of the building with (B) and (D) in Fig. 16: in the Northern and Southern part the delineation is quite accurate, demonstrating the influence of used straight lines. This is expected since the reference used for the ISPRS benchmark are captured from the same images.



REF→	Building	Seal_Grd
Building	**0.967**	0.033
Seal_Grd	0.101	**0.899**

Confusion matrix RTrees Area 5 using own reference data, overall accuracy 96.3%

REF→	Building	Seal_Grd
Building	**0.882**	0.118
Seal_Grd	0.004	**0.996**

Confusion matrix MRF Area 5 using own reference data, overall accuracy 90.2%

	Building
Completeness area	78.5
Correctness area	92.2
Completeness obj	81.6
Correctness obj	29.9
Completeness objXL	77.6
Correctness objXL	91.2
RMS [m]	1.1

Evaluation RTrees Area 5 using benchmark reference data

	Building
Completeness area	73.2
Correctness area	92.8
Completeness obj	76.3
Correctness obj	20.4
Completeness objXL	82.9
Correctness objXL	73.8
RMS [m]	1.4

Evaluation MRF Area 5 using benchmark reference data

Fig. 14. Results Area 5.

5. Discussion

On average the building detection can be considered successful; larger objects are quite completely and correctly detected by both strategies, supervised and unsupervised. Exceptions are large areas of the Toronto dataset and Vaihingen, area 2, where tree regions got labeled as buildings.

Vegetation detection, and discrimination between sealed and vegetated areas, rely on NDVI or RGB, but we saw that this is a problem in shadowed areas. It is even more critical if – like seen in the Toronto dataset – only RGB from a non-vegetation season is available. While in the supervised approach the effects are mitigated because shadowed areas can also be part of the training sample, the MRF technique will fail in those cases.

Another problematic issue concerns the height definition, especially for the differentiation between bushes and trees; are the former one trees as well, or is it low vegetation? For both classification strategies this fuzziness leads to misclassification,

Table 1
Results image-based point cloud densification for buildings.

	Area 1		Area 2		Area 3	
Compl. area	87.4	−3.8	88.2	−5.8	83.3	−5.8
Corr. area	95.2	+4.9	98.0	+9.0	95.8	+3.3
Compl. obj	83.8	−2.7	78.6	±0	73.2	−1.8
Corr. obj	100	+8.6	100	+57.7	97.7	+19.5
Compl. objXL	100	±0	100	±0	92.1	−2.6
Corr. objXL	100	±0	100	±0	100	±0
RMS (m)	1.0	−0.1	1.0	+0.2	1.1	+0.3

Evaluation of buildings using ISPRS benchmark, comparison to Random Trees results, Vaihingen.

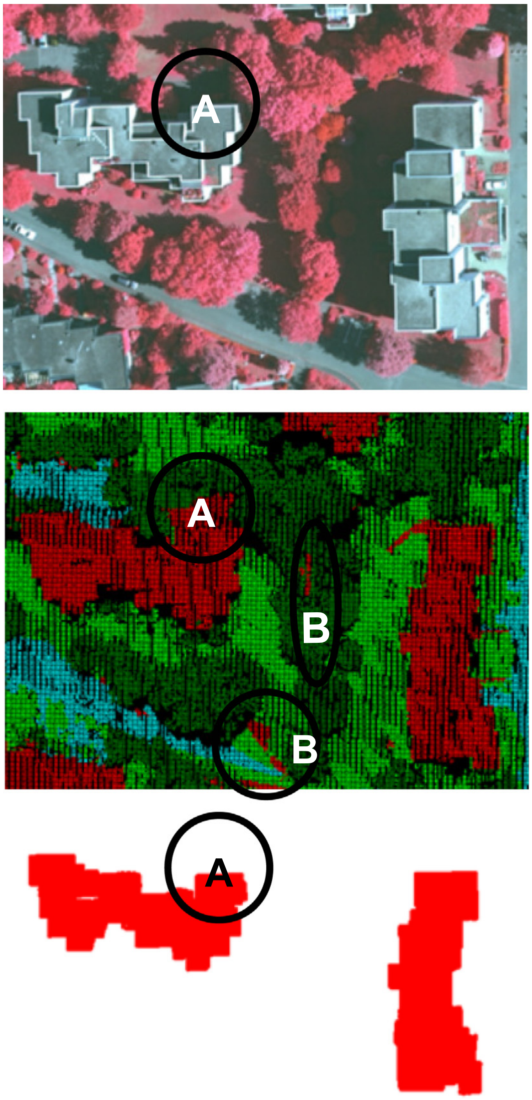


Fig. 15. Image-based point cloud densification, examples from area 2: upper image: false color, middle image: initial RTree classification, lower image: result of densification (buildings only). A: false positive in shadow, B: previous false positives got removed. For the colored figure please refer to the online version of this article.

but the supervised, similar as above, can adapt better to different situations. As far as the geometry of trees is concerned it plays a role whether in the reference tree crowns are modeled as circles (as done for the ISPRS reference) or as an individual area. This difference in object definition is the reason for some mismatch between the own evaluation and the ISPRS benchmark result.

We also saw another example for the problem of fuzzy class definition: based on some high level knowledge and experience a human operator labels objects. Without higher level reasoning some cases cannot be solved by automatic methods. For instance, concerning the ground-level attachment to the building in area 2 a method would need to detect walls, find out that in this case the wall is completely connected to the building and thus semantically declare this a part of the actual building.

The new segmentation technique which exploits both, geometry from ALS and spectral information, enhances the classification result in areas where different land cover is placed on the same physical plane. However, in case of undersegmentation, for example because of high vegetation adjacent to buildings, the classification might be wrong and the quality of the object outline is hampered, as well.

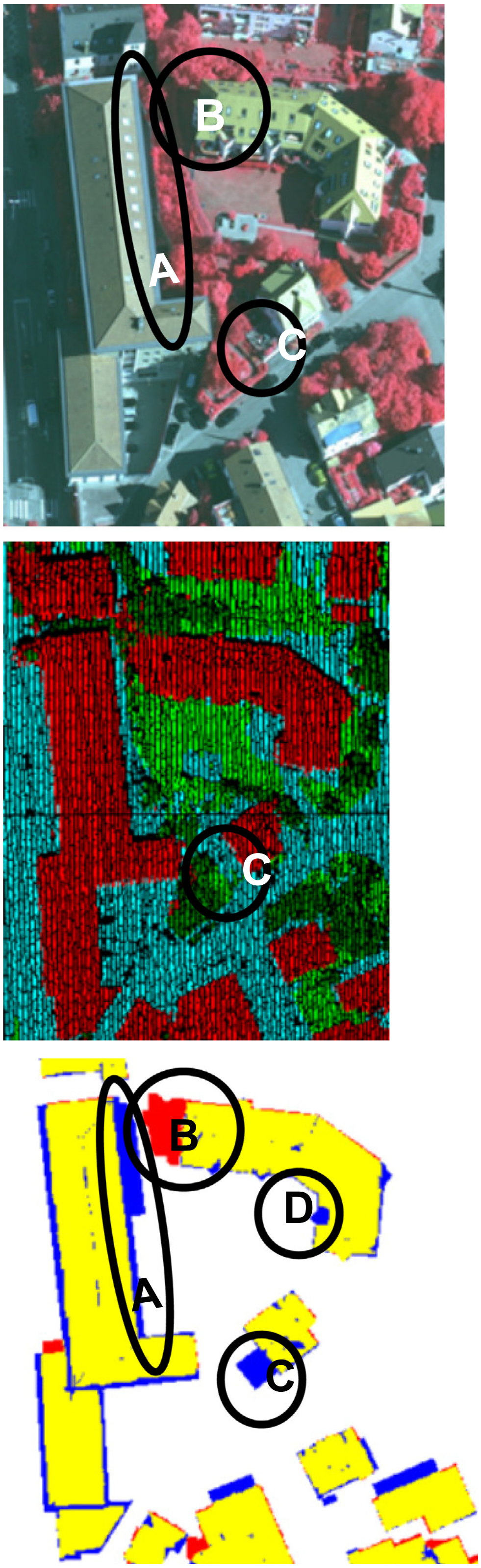


Fig. 16. Image-based point cloud densification, examples from area 1: upper image: false color, middle image: initial RTree classification, lower image: result from evaluation of densification (buildings only), yellow: true positive pixels, red: false positive, blue: false negative. (A) Missing plane, (B) false positive in shadow, (C) still missing building, (D) missing plane. For the colored figure please refer to the online version of this article.

Concerning the image-based point cloud densification for buildings we observed that shadow areas are a main problem here as well, since homogeneous texture is assumed an indicator for buildings, and in shadowed areas we might obtain oversegmentation. Further, in the current method we need to set a threshold for a minimum number of points located in an image segment. Such strict thresholds prevent smaller roof planes from consideration. Besides this, the majority of initial regions labeled wrongly as building got removed and the outline of correctly refined planes fits well to the reference.

6. Conclusions and outlook

In this paper, we present an approach which integrates ALS point cloud and image-based features in object space for 3D scene interpretation. The newly developed point cloud segmentation which exploits color for subdividing planes and point clusters works well in most cases. However, it turned out that the NDVI channel which is used for sub-segmentation is quite sensitive to shadow and this might lead to undersegmentation effects. The reliance on the NDVI also poses problems for the classification, especially the unsupervised MRF-based approach. For this reason it will be tested whether the brightness channel can be used in addition to include a proper weighting of the NDVI in shadowed areas. To further mitigate the impact of wrongly classified vegetation segments a shape indicator like compactness could be used to discriminate buildings and trees, also in the MRF formulation.

The MRF-based energy formulation is still not consequently segment-based; the individual voxels retrieve per class an energy derived from the segment it is assigned to, but further the segmentation information is not exploited. It would therefore be interesting to test to assign individual energy values per original voxel (and thus avoid the smoothing of individual feature values), but introduce a penalty for label changes across segment boundaries.

The object-space integration of geometric (point-based) and low to mid level feature information from images is a very useful approach for the interpretation of oblique airborne images. In those images façade parts are visible and a traditional 2.5D approach will not work, or leave out very important information, respectively. We showed already that the combination of point cloud and image information for rule-based façade detection from those images works out quite well (Xiao et al., 2012). To interpret the entire scene we modified the approach shown here towards façades, as well (Gerke and Xiao, 2013). More interesting would be to extend the neighborhood term from the MRF-based energy formulation towards a more semantic driven procedure. For instance, it is quite unlikely that façade voxels are above roof voxels and such configurations can be penalized in the model.

The two main problems with the image-based point cloud densification – reliance on color and the use of strict thresholds for points located in a segment – can be approached on the one hand similar as proposed above: by introducing brightness and the normalized height information in addition, and on the other hand by considering neighborhood relations. If a small plane is not isolated but adjacent to a larger plane it is likely to be part of the roof.

Acknowledgements

The Vaihingen data set was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) (Cramer, 2010). The authors would like to acknowledge the provision of the Downtown Toronto data set by Optech Inc., First Base Solutions Inc., GeoICT Lab at York University, and ISPRS WG III/4. Further, we made use of the implementation of graph-cut optimization by Boykov and Kolmogorov (2004). We thank the anonymous reviewers for their valuable comments.

References

Amhar, F., Josef, J., Ries, C., 1998. The generation of true orthophotos using a 3D building model in conjunction with a conventional dtm. In: ISPRS Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 32 (4), pp. 16–22.

Ardila Lopez, J.P., Tolpekin, V.A., Bijker, W., Stein, A., 2011. Markov-random-field-based super-resolution mapping for identification of urban trees in VHR images. ISPRS Journal of Photogrammetry and Remote Sensing 66 (6), 762–775.

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

Axelsson, P., 2000. DEM generation from laser scanner data using adaptive TIN models. In: ISPRS Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 23 (4), pp. 110–117.

Barnea, S., Filin, S., 2013. Segmentation of terrestrial laser scanning data using geometry and image information. ISPRS Journal of Photogrammetry and Remote Sensing 76, 33–48.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing 65 (1), 2–16.

Boykov, Y., Kolmogorov, V., 2004. An experimental comparison of min-cut/max-flow algorithms Fokr energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (9), 1124–1137.

Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (11), 1222–1239.

Brédif, M., Tournaire, O., Vallet, B., Champion, N., 2013. Extracting polygonal building footprints from digital surface models: a fully-automatic global optimization framework. ISPRS Journal of Photogrammetry and Remote Sensing 77, 57–65.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.

Bulatov, D., Rottensteiner, F., Schulz, K., 2012. Context-based urban terrain reconstruction from images and videos. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 1 (3), pp. 185–190.

Burns, J., Hanson, A., Riseman, E., 1986. Extracting straight lines. IEEE Transactions on Pattern Analysis and Machine Intelligence 8 (4), 425–455.

Comaniciu, D., Meer, P., 2002. Mean shift: a robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (5), 603–619.

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

Felzenszwalb, P., Huttenlocher, D., 2004. Efficient graph-based image segmentation. International Journal of Computer Vision 59 (2), 167–181.

Freund, Y., Schapire, R.E., 1996. Experiments with a new boosting algorithm. In: Proceedings of the Thirteenth International Conference of Machine Learning. Morgan Kaufman, Bari, Italy, pp. 148–156.

Gée, C., Bossu, J., Jones, G., Truchetet, F., 2008. Crop/weed discrimination in perspective agronomic images. Computers and Electronics in Agriculture 60 (1), 49–59.

Gerke, M., 2011. Supervised classification of multiple view images in object space for seismic damage assessment. Lecture Notes in Computer Science: Photogrammetric Image Analysis Conference 2011, vol. 6952. Springer, pp. 221–232.

Gerke, M., Xiao, J., 2013. Supervised and unsupervised MRF based 3D scene classification in multiple view airborne oblique images. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 2 (3) pp. 25–30.

Gordon, J., Shortliffe, E.H., 1990. The Dempster–Shafer theory of evidence. In: Shafer, G., Pearl, J. (Eds.), Readings in Uncertain Reasoning. Morgan Kaufmann, San Mateo, CA, pp. 529–539.

Grigillo, D., Kanjir, U., 2012. Urban object extraction from digital surface model and digital aerial images. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 1 (3), pp. 215–220.

ISPRS WG III/4, 2013. Homepage of ISPRS working group III/4 "3D scene analysis". <http://www2.isprs.org/commissions/comm3/wg4.html>. (accessed 15.06.13).

Katz, S., Tal, A., Basri, R., 2007. Direct visibility of point sets. ACM Transactions on Graphics 26 (3), 24.

Khoshelham, K., Nardinocchi, C., Frontoni, E., Mancini, A., Zingaretti, P., 2010. Performance evaluation of automated approaches to building detection in multi-source aerial data. ISPRS Journal of Photogrammetry and Remote Sensing 65 (1), 123–133.

Kim, C., Habib, A., 2009. Object-based integration of photogrammetric and Lidar data for automated generation of complex polyhedral building models. Sensors 9 (7), 5679–5701.

Lafarge, F., Mallet, C., 2012. Creating large-scale city models from 3D-point clouds: a robust approach with hybrid representation. International Journal of Computer Vision 99 (1), 69–85.

Li, S.Z., 2009. Markov Random Field Modeling in Image Analysis. Advances in Computer Vision and Pattern Recognition third ed. Springer, Berlin.

Niemeyer, J., Rottensteiner, F., Sörgel, U., 2012. Conditional random fields for Lidar point cloud classification in complex urban areas. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 1 (3), pp. 263–268.

Nyaruhuma, A.P., Gerke, M., Vosselman, G., Mtalo, E.G., 2012. Verification of 2D building outlines using oblique airborne images. ISPRS Journal of Photogrammetry and Remote Sensing 71, 62–75.

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

Rapidlasso, 2013. Homepage lastools. <http://lastools.org> (accessed 15.06.13).

Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., Breitkopf, U., 2012. ISPRS benchmark on urban object classification and 3D building reconstruction. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 1 (3), pp. 293–298.

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

Sampath, A., Sha, J., 2007. Building boundary tracing and regularization from airborne lidar point clouds. Photogrammetric Engineering & Remote Sensing 73 (7), 805–812.

Sampath, A., Shan, J., 2010. Segmentation and reconstruction of polyhedral building roofs from aerial lidar point clouds. IEEE Transactions on Geoscience and Remote Sensing 48 (3), 1554–1567.

Vosselman, G., Gorte, B., Sithole, G., Rabani, T., 2004. Recognising structure in laser scanner point clouds. In: ISPRS Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 36 (8), pp. 33–38.

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. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 1–7, pp. 323–328.

Xiao, J., 2012. Automatic building outlining from multi-view oblique images. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 1 (3), pp. 323–328.

Xiao, J., Gerke, M., Vosselman, G., 2012. Building extraction from oblique airborne imagery based on robust facade detection. ISPRS Journal of Photogrammetry and Remote Sensing 68, 56–68.

Xu, S., Oude Elberink, S., Vosselman, G., 2012. Entities and features for classification of airborne laser scanning data in urban area. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 1 (4), pp. 257–262.

Zebedin, L., Klaus, A., Gruber-Geymayer, B., Karner, K., 2006. Towards 3D map generation from digital aerial images. ISPRS Journal of Photogrammetry and Remote Sensing 60 (6), 413–427.