Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Fusion 6 (2005) 283–300
INFORMATION FUSION
www.elsevier.com/locate/inffus

Using the Dempster–Shafer method for the fusion of LIDAR data and multi-spectral images for building detection

Franz Rottensteiner [a,*], John Trinder [a], Simon Clode [b], Kurt Kubik [b]

[a] School of Surveying and Spatial Information Systems, The University of New South Wales, Sydney, NSW 2052, Australia
[b] The Intelligent Real-Time Imaging and Sensing Group, School of ITEE, The University of Queensland, Brisbane QLD 4072, Australia

Received 30 September 2003; received in revised form 16 June 2004; accepted 17 June 2004
Available online 20 July 2004

Abstract

A method for the detection of buildings in densely built-up urban areas by the fusion of first and last pulse laser scanner data and multi-spectral images is presented. The method attempts to achieve a classification of land cover into the classes "building", "tree", "grassland", and "bare soil", the latter three being considered relevant for the subsequent generation of a high-quality digital terrain model (DTM). Building detection is accomplished by first applying a hierarchical rule-based technique for coarse DTM generation based on morphological filtering. After that, data fusion based on the theory of Dempster–Shafer is used at two different stages of the classification process. We describe the algorithms involved, giving examples for a test site in Fairfield (New South Wales).

Keywords: Airborne laser-scanning; Building detection; Dempster–Shafer; Classification; 3D city models

1. Introduction

1.1. Motivation and goals

Automation in data acquisition for 3D city models is an important topic of research in photogrammetry. In addition to aerial images, point clouds generated by airborne laser scanning, also known as LIDAR (*LI*ght *D*etection *A*nd *R*anging), are being used for that purpose more frequently. This development has been triggered by the progress in sensor technology, which has rendered possible the acquisition of very dense point clouds using airborne laser scanners [1]. Problems to be tackled in this context comprise the generation of high-quality digital terrain models (DTMs) in urban areas and the extraction of topographic objects such as buildings or trees. These problems are closely interrelated. For computing a DTM, the LIDAR points on the tops of buildings and trees have to be eliminated, and thus information about the positions of such objects is required, whereas on the other hand, a DTM is required if buildings or trees are to be detected. This means that the first step to be carried out in order to extract meaningful semantic information from LIDAR data is a classification of data to separate points situated on the terrain from those on buildings, trees, and other objects. With high-resolution LIDAR data, it is not only possible to perform this classification, but also to reconstruct objects such as buildings geometrically at a relatively high level of detail [2–4].

With the decreasing resolution of the LIDAR data, the classification becomes more difficult, especially in residential areas characterised by detached houses, where the appearance of trees and buildings in the data might be similar. In order to improve the classification results, additional data sources can be considered. First, LIDAR systems register two echoes of the laser beam, the *first* and the *last pulse*, corresponding to the first and the last points from where the laser beam is reflected. Differences between the first and last pulse data indicate

* Corresponding author. Tel.: +61-2-9385-4186; fax: +61-2-9313-7493.
E-mail addresses: f.rottensteiner@unsw.edu.au (F. Rottensteiner), j.trinder@unsw.edu.au (J. Trinder), sclode@itee.uq.edu.au (S. Clode), kubik@itee.uq.edu.au (K. Kubik).

1566-2535/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.inffus.2004.06.004

height steps, e.g. at building boundaries, power lines, and, most frequently, trees [1]. Second, the *intensity* of the returned laser beam can be used. LIDAR systems typically operate in the infrared part of the electromagnetic spectrum. However, these data have the disadvantage that they are under-sampled and, thus, very noisy [5], because the footprint size of the laser beam is typically 0.2 m, while the average point distance is 1 m. Third, *multi-spectral images* can provide valuable information due to their spectral content and because their resolution is better than the resolution of laser scanner data [6].

It is the goal of this paper to investigate the fusion of first and last pulse LIDAR data and multi-spectral images for building detection in densely built-up urban areas. In the course of the building detection process, fusion will be especially helpful for improving the classification of land cover. Although our emphasis is on the detection of buildings, we also want to distinguish three other classes that are relevant in the context of 3D city models and for the generation of high-quality DTMs in urban areas: trees, grassland, and bare soil. This is accomplished by first applying a hierarchical technique for coarse DTM generation, followed by the fusion of various cues derived from the input data for classification. Here, data fusion is based on the theory of Dempster–Shafer [7]. Examples are presented for a test site in Fairfield (New South Wales) covering an area of 2×2 km^2.

1.2. Background

Existing classification techniques for building detection from LIDAR and/or multi-spectral data can be characterised by the cues that are actually used for classification and by the methods used for the data fusion. These topics will be discussed separately.

1.2.1. Cues for building detection

From the LIDAR data, a *digital surface model* (DSM) can be derived, for example by sampling the data into a regular grid [3,8]. The DSM represents the surface from which the laser pulse is reflected, that is trees, terrain surface, buildings, etc. In flat terrain, the elevations of the DSM can be used directly to separate elevated objects from others. In the case of undulating terrain, the elevations reflect both the terrain heights and the height differences between points on elevated objects and the terrain. That is why the classification of LIDAR data for the automatic detection of topographic objects usually starts with the generation of a DTM, e.g. by morphological opening filters or rank filters [9] or by hierarchical robust linear prediction [10]. We apply a hierarchical method of morphological filtering for DTM generation in this study [11], which will be outlined in Section 2.3. The DTM is subtracted from the DSM, yielding a *'normalised' DSM*, the heights of which directly reflect the heights of objects relative to the terrain [9].

A DSM also provides information about *local surface properties* via an analysis of the derivatives of the DSM. The *maximum slope* has been used for distinguishing flat roofs from tilted roofs [8]. The second derivatives are more commonly used for building detection. As they are closely related to the local curvature of the DSM, they can be used to derive measures for *surface roughness*. Assuming that roofs mostly consist of planar or at least smooth surfaces, an analysis of surface roughness can help to separate buildings from trees. Various parameters have been used in the past to characterise surface roughness, e.g. the output of a Laplace filter applied to the DSM [8,12], local curvature [12], or the local variance of the surface normal vectors [13]. The model for surface roughness we use is based on an analysis of the local variations of the surface normal vectors using the framework of polymorphic feature extraction [14]. It will be described in Section 2.2.

Height differences between first and last pulse data have also been used to improve the results of building detection. They can be used as an indicator of vegetation, but large differences also occur at the edges of buildings. Until recently, laser scanners only could deliver either first or last pulse data, and the user of the data had to select one of these scanning modes according to his or her priorities. Modern laser scanners can deliver both first and last pulse data in the same flight. A detailed discussion of the effects of choosing one of the two scanning modes on the results of building detection is given in [15]. In addition to the local characteristics described up to now, *average surface roughness parameters* or the *size of building candidate regions* can be evaluated to eliminate candidate regions being either oddly shaped, too large, too small, or characterised by a large average surface roughness [3,9].

Schenk and Csatho [6] put forward the idea of exploiting the complementary properties of LIDAR data and *aerial imagery* to first achieve a more complete surface description by a feature based fusion process and then to extract semantically meaningful information from the aggregated data. They pointed out that LIDAR data are especially useful for the detection of surface patches having specific geometrical properties, and for deriving other properties such as their roughness. On the other hand, aerial images can help to provide the surface boundaries and the locations of surface discontinuities.

Haala and Brenner [2] combine a normalised DSM from LIDAR data with the three spectral bands of a scanned *colour infrared (CIR) image*. As the separation of trees from buildings is the most problematic task in this context due to the relatively low resolution of the LIDAR data, it is important to include the near-infrared

band in the classification process. As an alternative to using all bands of multi-spectral images, the *normalised difference vegetation index (NDVI)* derived from the near infrared and the red portions of the spectrum can be applied for its potential in discriminating vegetation [16].

1.2.2. Data fusion techniques for building detection

Building detection is usually carried out in two stages:

1. *Detection of building candidate segments:* An initial classification is carried out on a per-pixel level. Each pixel is classified according to whether it is a building candidate pixel or not. The simplest way of doing so is applying a height threshold to the normalised DSM [9], but more sophisticated classification techniques can be applied, too. Connected components of building pixels then become building candidate regions.
2. *Evaluation of the building candidate regions:* For the initial regions, average parameters describing the DSM heights, the spectral characteristics, surface roughness, and the region size are also evaluated to separate the buildings from the trees [3,9,16]. Again, various techniques can be applied to combine the available cues.

The results of an object-based classification, which is applied at the second stage above, cannot be better than the results of the initial segmentation. For instance, if trees standing close to buildings are merged in the initial segmentation, they can no longer be separated. In [3], we have shown how the segmentation results can be improved at a later processing stage, but in general it would be preferable to improve the initial segmentation results by applying better fusion techniques in the first stage of classification. Some post-processing will also be required in that case: per-pixel classifiers do not consider the local context of a pixel or relationships with neighbouring pixels, thus there will be isolated spurious pixels. The classification results are especially unreliable at the building boundaries [8].

The simplest way to combine the different cues at each pixel is to concatenate the data from each cue to form a multi-dimensional data vector. The resultant vector can be treated as if it were from a single source [17]. That data vector can be used in standard procedures such as supervised maximum likelihood (ML) classification, unsupervised classification, or rule-based classification. As an alternative, Le Hégarat-Mascle et al. [18] name three more sophisticated techniques: fuzzy logic, probabilistic reasoning, and the theory of Dempster–Shafer. In the remainder of this section we want to give an overview of how these techniques have previously been used for building detection.

Supervised ML classification is applied to the DSM heights, slopes, and the results of a Laplace filter by Maas [8]. He uses sparse training regions containing the object classes 'flat roof', 'tilted roof', 'vegetation', 'flat terrain', and 'no data', but states that the selection of training regions might be replaced by introducing a priori knowledge about the class centres and covariance matrices to replace the interactive selection of training regions. However, it has already been pointed out in [17] that this may not be suitable when the various sources cannot be described by a common "spectral" model, especially when spectral and elevation data are combined. For instance, neither the relative heights nor the spectral characteristics of buildings can be assumed to be normally distributed. Buildings have different heights and spectral "colours", so that they correspond to more than one cluster in feature space.

Unsupervised classification based on the ISODATA algorithm is applied to the three bands of a CIR image and a normalised DSM by Haala and Brenner [2]. They stress the importance of the elevation data for the separability of the feature classes. The interpretation of the detected feature clusters is performed interactively. Appropriate modelling of the relevant classes to perform that interpretation automatically seems to be difficult. Lu and Trinder [16] applied a K-means algorithm to RGB images to obtain an initial segmentation and evaluate additional sources such as the NDVI and a DSM derived from image matching to automatically assign the feature clusters to thematic classes. The main problem with their approach is that the DSM is not used in the original unsupervised classification, so that shadows cause relevant building parts to be missed entirely.

Rule-based classification is based on expert knowledge about the appearance of certain object classes in the data that are used to define rules by which the classes can be separated. These rules often involve thresholding operations, as in the method we will outline in Section 2.3. Selecting "hard" thresholds properly is a critical issue. In addition, the hierarchical structure of many rule-based approaches, where first a subset of the cues is selected to make an initial classification and then the other cues are used to resolve any ambiguities [17], makes it impossible to recover from previous errors in the classification process. For this reason, we believe that algorithms evaluating all cues simultaneously are preferred, and even more so if the reliability of the results can be quantified.

Fuzzy logic can be used to model vague knowledge about class assignment in order to avoid hard thresholds as in rule-based algorithms [19]. In [12], the membership functions for a fuzzy logic classification using various cues derived from first and last pulse LIDAR data are described. These membership functions had to be defined for every class in a training phase. In a second step, these membership values were combined to obtain a final decision. This fuzzy classification is applied to

previously detected building candidate segments, rather than on a per-pixel basis, and the method was used in a supervised context.

Probabilistic reasoning aims at assigning each pixel to the class $C_i \in \{C_1, \ldots, C_N\}$ maximising the a posteriori conditional probability $P(C_i|\mathbf{x_s})$ of C_i given the data vector $\mathbf{x_s}$ for a pixel or feature s. The conditional probabilities are computed using the theorem of Bayes [20] which in turn requires the modelling of the a priori probabilities $P(\mathbf{x_s}|C_i)$ of data vector $\mathbf{x_s}$ under the assumption of a class C_i and $P(C_i)$ of class C_i. The probabilities $P(\mathbf{x_s}|C_i)$ are often modelled by a multivariate Gaussian distribution. Initially, the prior $P(C_i)$ is often assumed to be equal for all classes, and then iteratively recomputed from the relative numbers of pixels in the first iteration. As stated previously, modelling of these a priori probabilities becomes difficult if no training samples are used, especially if the assumption of a normal distribution of the data vectors is unrealistic, e.g. for built-up areas [20]. An example of combining various cues in a Bayesian network with the goal of detecting buildings was presented in [13]. A hierarchical strategy is used, turning the classification results of the coarser resolution into one of the cues for the classification in the next iteration. The model of the conditional probabilities relating the classification results is heuristic, which reduces the statistical soundness of the probabilistic approach. We also doubt that the DSM heights of the building roofs are normally distributed, but rather expect a mixture of several normal distributions, each corresponding to a specific building type.

Dempster–Shafer theory of evidence was introduced as an expansion of the probabilistic approach that can also handle imprecise and incomplete information as well as conflict within the data [7,17,18]. A description of the advantages of the Dempster–Shafer theory, compared to probabilistic reasoning, is given in [18]. An important property of that theory is its capability to handle the union of classes. In [18], it was applied to unsupervised classification of optical and SAR images. In the context of building detection, the Dempster–Shafer theory has been applied for the final classification of building candidate regions, combining cues such as the NDVI and the average relative heights to distinguish buildings from other objects [16].

Even though several authors assess the advantages of the theory of Dempster–Shafer for data fusion in classification [7,17,18], to our knowledge, that theory has not yet been applied to a per-pixel classification of high-resolution remotely sensed data of different origin with the goal of detecting individual buildings. In this paper, we will show how this can be accomplished for the fusion of LIDAR data and multi-spectral images without using training areas or assigning thematic labels to classes of feature space interactively. The Dempster–Shafer theory is applied in two stages of the overall process, first to detect the candidate building regions and then to eliminate false building candidates. We will describe a heuristic model for the distribution of the evidence provided by our cues to the classes of the classification problem, exploiting the fact that the Dempster–Shafer theory can handle the union of classes, which, in accordance with [18], we consider to be a good alternative to other classification techniques for handling the "mixed pixel" problem. We will also evaluate our method using reference data, showing that our method gives satisfactory results in an area of very inhomogeneous building shapes, and that the detectability of buildings mainly depends on the building size, given the resolution of the LIDAR data.

2. Fusing LIDAR data and multi-spectral images for building detection

2.1. Overview of the process flow

The input to our method is given by three data sets that have to be generated from the raw data by preprocessing. The *DSM corresponding to the last pulse data* ($\mathrm{DSM_L}$) is a regular height grid interpolated by linear prediction using a straight line as the covariance function, thus almost without filtering. This is accomplished by using the program SCOP developed at Vienna University of Technology [3]. The first pulse data are also sampled into a regular grid, and by computing the height differences of the first and the last pulse DSMs, a grid ΔH_{FL} of the *height differences between the first and the last pulses* is obtained. The NDVI is computed from the near infrared and the red bands of the multi-spectral images [16]. The image data must be geocoded so that the data are already aligned for the subsequent processes.

The work flow for building detection consists of two stages. First, a coarse DTM is generated from the input data. This DTM is used to compute a normalised DSM, which along with parameters of surface roughness of the $\mathrm{DSM_L}$, the NDVI, and the height differences ΔH_{FL}, provides the input for the second stage, the detection of building regions by Dempster–Shafer fusion. As the model we use for surface roughness is an important component of our method, it will be described in Section 2.2.

Coarse DTM generation is based on the hierarchical application of morphological grey scale opening using structural elements of different sizes to overcome the problems caused by large buildings in the data set. After morphological opening, a rule-based algorithm is used to detect large buildings in the data. That information is used in the next iteration, when a smaller structural element is used for morphological opening, to eliminate

large buildings. DTM heights computed from the previous iteration are substituted for the results of the morphological filter. The process is finished when the minimum size for the structural element is reached. Our method for hierarchical DTM generation is described in more detail in Section 2.3. Note that the classification performed here was originally used for building detection alone [11]. As this rule-based algorithm consists of a sequence of thresholding operations, all the available information is never evaluated jointly. Classification errors committed in one of the thresholding operations cannot be corrected in the subsequent stages of the algorithm. Thus, we use this algorithm only to eliminate large buildings in DTM generation (when it is also simple to select some of the thresholds because large buildings are usually characterised by large roof planes). The more sophisticated Dempster–Shafer fusion technique is then used in the final stages of building detection.

The work flow for building detection based on Dempster–Shafer fusion is presented in Fig. 1. First, the normalised DSM (nDSM) and two parameters describing the roughness of the DSM_L, namely the strength and the directedness of surface roughness (cf. Section 2.2) are computed. Altogether, there are five data sets that contribute to a Dempster–Shafer fusion process carried out for each pixel of the DSM_L grid independently. Each pixel is assigned to one of four classes, namely *building* (B), *tree* (T), *grassland* (G), and *bare soil* (S). In the subsequent steps, the binary image of the building pixels is used. Morphological filtering helps to eliminate small areas of pixels erroneously classified as building pixels. Then, connected components of building pixels are sought, which results in initial building regions. As the first fusion step accounts for only a very small local neighbourhood (for the evaluation of surface roughness), we eliminate spurious initial building regions in a second Dempster–Shafer fusion process which utilizes the average NDVI, the average nDSM heights, and two additional attributes derived from the surface roughness parameters. Building detection by data fusion based on the theory of Dempster–Shafer is described in Section 2.4.

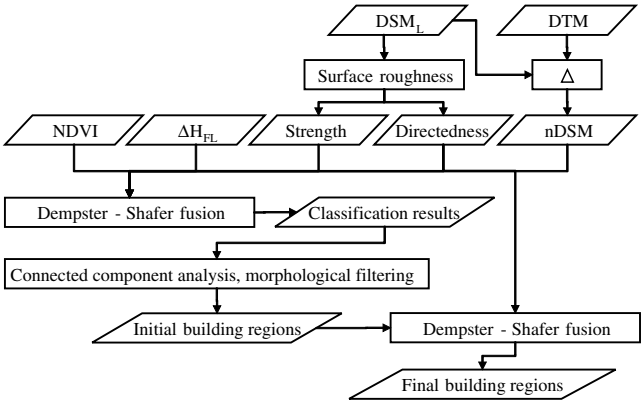

Fig. 1. The work flow for building detection by data fusion based on the theory of Dempster–Shafer.

2.2. Surface roughness

In this section, the model for DSM surface roughness, which is based on the framework of polymorphic feature extraction [14,21], is presented. In polymorphic feature extraction, the digital image is assumed to consist of regions of homogeneous grey level vectors ("segments"), line regions, and point regions, the latter two being a result of blurring effects in the sensor. The grey level vectors $\mathbf{g}(x,y) = [g_1(x,y), \ldots, g_K(x,y)]^T$ are sampled in a grid (x,y); K represents the number of bands of the image. The grey levels $g_i(x,y)$ of band i are modelled assuming they are affected by additive noise $n_i(x,y)$. For digital images, the noise is assumed to be Poisson-distributed, but the distribution is approximated by a normal distribution $N(0,\sigma_{ni}^2)$ with a signal-dependent noise variance σ_{ni}^2 [21]. For our application, σ_{ni}^2 corresponds to the variance of height differences, which can be assumed to be Gaussian.

The first task in polymorphic feature extraction is the classification of each pixel according to whether it is situated in a homogeneous region, a line region, or a point region. With Δg_{ix} and Δg_{iy} denoting the first derivatives of the grey levels of band i by x and y respectively, a matrix $\mathbf{N}$ is computed:

$$\mathbf{N} = \sum_{i=1}^{K} \frac{1}{\bar{\sigma}_{ni}^2} \cdot \mathbf{L} * \begin{pmatrix} \Delta g_{ix}^2 & \Delta g_{ix} \cdot \Delta g_{iy} \\ \Delta g_{ix} \cdot \Delta g_{iy} & \Delta g_{iy}^2 \end{pmatrix} \tag{1}$$

In Eq. (1), $\mathbf{L}$ is a lowpass filter by which the elements of the matrix are convolved. $\bar{\sigma}_{ni}^2$ is the variance of the smoothed grey level differences $\Delta\bar{g}_{ix} = \mathbf{L} * \Delta g_{ix}$ and $\Delta\bar{g}_{iy} = \mathbf{L} * \Delta g_{iy}$, which can be derived from σ_{ni}^2 by error propagation [14]. Using the elements of $\mathbf{N}$ and denoting the eigenvalues of $\mathbf{N}$ by λ_1 and λ_2, a measure R of homogeneity or *texture strength* and a measure D for *texture directedness* can be defined [14,21]:

$$R = \mathrm{tr}(\mathbf{N}) = \sum_{i=1}^{K} \frac{\mathbf{L} * (\Delta g_{ix}^2 + \Delta g_{iy}^2)}{\bar{\sigma}_{ni}^2} \tag{2}$$

$$D = 1 - \left(\frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2}\right)^2 = \frac{4 \cdot \det(\mathbf{N})}{\mathrm{tr}^2(\mathbf{N})} \tag{3}$$

Using these two measures, the classification of each pixel is performed. Using a threshold $R_{\min}$ for texture strength, a pixel is classified as being in a homogeneous region if $R \leqslant R_{\min}$. If $R > R_{\min}$, the local neighbourhood (defined by the extents of the filter $\mathbf{L}$) contains significant grey level variations, i.e., variations of a size that can no longer be explained by noise with a variance σ_{ni}^2. Texture directedness is used to decide whether these

variations are isotropic or not. Using a threshold $D_{\min}$, the distribution of the directions of the gradient vectors is considered to be isotropic if $D > D_{\min}$, and the pixel is classified as being in a point-like region. If $D \leqslant D_{\min}$, the gradient vectors are supposed to be more or less parallel, which indicates that the pixel belongs to a line region.

The selection of the threshold $R_{\min}$ is very critical. If good estimates for σ_{ni}^2 are available, e.g. derived in the way described in [21], R can be assumed to follow a χ^2 distribution, and $R_{\min}$ is selected relative to the significance level of a hypotheses test [14]. In some occasions it might be convenient to select

$$R_{\min} = j \cdot \text{median}(R) \tag{4}$$

By relating $R_{\min}$ to the median of R, the selection of that threshold is replaced by the selection of a multiplication constant j. Again, this is not too critical because $R_{\min}$ is adaptive to the image content (Eq. (4)). The selection of the threshold $D_{\min}$ for texture directedness is less critical because D is always between 0 and 1. $D_{\min}$ can be chosen to be between 0.5 and 0.7 [14].

We apply the classification scheme of polymorphic feature extraction to the first derivatives of the DSM. Assuming the DSM to be represented by a height grid $z(x,y)$, we obtain a digital "image" of two bands. The grey levels of this image are the first derivatives of $z(x,y)$ by x and y, $\mathbf{g}(x,y) = (\partial z/\partial x, \partial z/\partial y)^{\mathrm{T}}$ respectively. Under these assumptions, the matrix $\mathbf{N}$ can be computed from:

$$\mathbf{N} = \frac{1}{\bar{\sigma}_x^2} \cdot \mathbf{L} * \begin{pmatrix} \left(\frac{\partial^2 z}{\partial x^2}\right)^2 & \left(\frac{\partial^2 z}{\partial x^2}\right) \cdot \left(\frac{\partial^2 z}{\partial x \partial y}\right) \\ \left(\frac{\partial^2 z}{\partial x^2}\right) \cdot \left(\frac{\partial^2 z}{\partial x \partial y}\right) & \left(\frac{\partial^2 z}{\partial x \partial y}\right)^2 \end{pmatrix} + \frac{1}{\bar{\sigma}_y^2} \cdot \mathbf{L} * \begin{pmatrix} \left(\frac{\partial^2 z}{\partial y \partial x}\right)^2 & \left(\frac{\partial^2 z}{\partial y \partial x}\right) \cdot \left(\frac{\partial^2 z}{\partial y^2}\right) \\ \left(\frac{\partial^2 z}{\partial y \partial x}\right) \cdot \left(\frac{\partial^2 z}{\partial y^2}\right) & \left(\frac{\partial^2 z}{\partial y^2}\right)^2 \end{pmatrix} \tag{5}$$

In Eq. (5), $\bar{\sigma}_x^2$ and $\bar{\sigma}_y^2$ are the variances of the smoothed matrix elements. They can be derived from an estimate of the variance σ_z^2 of the DSM heights by error propagation. The "grey levels" of the original image are the components of the surface normal vectors. The elements of $\mathbf{N}$ are derived from the second derivatives of the DSM and thus related to the local curvature of the DSM. "Homogeneous" pixels are situated in neighbourhoods of homogeneous surface normal vectors and thus in neighbourhoods of small second derivatives (Eqs. (2) and (5)). Texture strength R can thus be interpreted as a measure of local co-planarity, or as the strength of local *surface roughness*. "Line" pixels are situated in neighbourhoods with large, but anisotropic variations of the surface normal vectors. They correspond to surface discontinuities and surface intersection curves. Point pixels occur in areas with large amplitude, isotropic variations of the surface normal vectors, which is typical for building corners and for trees. In the subsequent sections, we will see how texture strength and texture directedness, as well as the results of texture classification are used as cues for building detection.

In our implementation, we choose $\mathbf{L}$ to be a Binomial filter kernel of size $n \times n$ and select $R_{\min}$ according to Eq. (4). Thus, there are three control parameters: the size n of the filter kernel, describing the extent of the neighbourhood for which texture classification is carried out, the multiplication constant j from Eq. (4), and $D_{\min}$. We choose n in accordance with the minimum linear extent of a roof plane we expect to be detectable, given the resolution of the DSM (e.g., $n = 3$). $D_{\min}$ is chosen to be 0.7. The most critical choice is the value of j. Typically, we select j to be between 1.0 and 2.0.

2.3. Hierarchical DTM generation

In this section, the process of DTM generation is discussed in detail. For building detection, the DTM should be a relatively good approximation of the terrain, so that the nDSM reflects the actual building heights. If the DTM is to be created by morphological filtering, the size of the structural element must reflect the size of the largest building available in the data set. That is, large structural elements are required in the presence of large buildings. However, this means that terrain structures smaller than the structural element, as well as large buildings, will be eliminated by morphological filtering. Wherever this occurs, the heights of the nDSM will be systematically distorted, because by "cutting off" the tops of small hills the heights of the nDSM may become too large. This will cause errors in classification algorithms when the nDSM is taken as an input. As morphological filtering using a structural element smaller than the largest buildings in the data set will not filter out these buildings, they must be detected and then eliminated from the DTM in the hierarchical procedure outlined earlier in Section 2.1. In each of the iterations, initial candidate regions for large buildings are detected by thresholding operations, and then surface roughness is evaluated to eliminate regions corresponding to trees.

2.3.1. Detecting candidate regions for large buildings

The morphological filtering provides a coarse approximation for the DTM. In all iterations except the first one, the DTM heights generated in the previous iteration are substituted for the results of morphological filtering in areas where buildings have been detected. This substitution ensures that large buildings that are preserved by morphological filtering are eliminated in the current iteration. Hence, an initial building mask is created by thresholding the heights of the nDSM. This initial building mask still contains singular pixels with large nDSM heights, areas covered by vegetation, and terrain structures smaller than the structural element for morphological filtering. Pixels having an NDVI greater

than a certain threshold or a height difference ΔH_{FL} greater than an appropriate threshold are considered to be covered by vegetation and thus erased from the building mask. Also, a binary morphological opening filter using a small structural element is applied to the initial building mask to remove oddly shaped objects and to separate building regions just bridged by a thin line of pixels. The initial building regions are obtained by a connected component analysis of the resulting image. Small regions are discarded, because at this stage, we only want to detect large buildings in order to eliminate them from the DTM.

2.3.2. Classification of building candidate regions

Some of the initial building regions correspond to groups of trees or to small terrain structures. These regions can be eliminated by evaluating a surface roughness criterion derived by an analysis of the second derivatives of the DSM_L, using the method described in Section 2.2. The numbers of "homogeneous" and "point-like" pixels are counted in each initial building region. Buildings are characterised by a large percentage of "homogeneous" and by a small percentage of "point-like" pixels. By comparing these percentages to thresholds, non-building regions can be eliminated. The surface roughness criterion performs well for large buildings and with dense LIDAR data [3]. However, if the point distance of the LIDAR data is larger, e.g. 1 m, only a few LIDAR points are situated on small buildings, so that the percentage of "homogeneous" pixels is reduced, while the percentage of "point-like" pixels is increased, causing the detection of small buildings to be more difficult.

2.4. Building detection based on Dempster–Shafer fusion

The Dempster–Shafer theory of evidence is frequently applied for the fusion of data from multiple sensors. Unlike Bayesian probabilistic reasoning, it offers tools to represent partial knowledge about a sensor's contribution to the classification process. We provide an overview on that theory based on [7,17,18] in Section 2.4.1. In Section 2.4.2, we present some general considerations with respect to the definition of the probability masses. Then, we describe the application of the Dempster–Shafer theory in building detection in Sections 2.4.3 and 2.4.4. Note that in Section 2.4.1, we use the term "sensor" in the way it is done in the cited literature, whereas in the other sections, we rather use the term "cue", because some of the cues we use are derived from one sensor only.

2.4.1. Theory of Dempster–Shafer fusion

Let us assume a classification problem where the input data are to be classified into n mutually exclusive classes. The set Θ of these classes is called the frame of discernment. The power set of Θ is denoted by 2^Θ. It contains both the classes and all their possible unions. In the theory of Dempster and Shafer, a probability mass $m(A)$ is assigned to every class $A \in 2^\Theta$ by a sensor (a cue for classification) such that $0 \leqslant m(A) \leqslant 1$, $m(\emptyset) = 0$, and $\sum_{A \in 2^\Theta} m(A) = 1$, with $\emptyset$ denoting the empty set.

Imprecision of knowledge can be handled by assigning a non-zero probability mass to the union of two or more classes. Two parameters, support $\mathrm{Sup}(A)$ and plausibility $\mathrm{Pls}(A)$, can be defined for all $A \in 2^\Theta$:

$$\mathrm{Sup}(A) = \sum_{B_S \subseteq A} m(B_S) \tag{6}$$

$$\mathrm{Pls}(A) = \sum_{A \cap B_{Pl} \neq \emptyset} m(B_{Pl}) = 1 - \mathrm{Sup}(\overline{A}) \tag{7}$$

In Eq. (6), the sum is taken over all classes $B_S \in 2^\Theta$ with $B_S \subseteq A$. The sum in Eq. (7) is taken over all $B_{Pl} \in 2^\Theta$ with $A \cap B_{Pl} \neq \emptyset$. The support of a class is the sum of all masses directly assigned to that class by a data source, whereas the plausibility sums all masses not assigned to the complement of a class. An uncertainty interval $[\mathrm{Sup}(A), \mathrm{Pls}(A)]$ with $\mathrm{Sup}(A) \leqslant \mathrm{Pls}(A)$ can be defined; its length is a measure of the imprecision of knowledge about the uncertainty of class A [18]. $\overline{A}$ is the complementary hypothesis of A: $A \cup \overline{A} = \Theta$ and $A \cap \overline{A} = \emptyset$. Its support $\mathrm{Sup}(\overline{A})$ represents the degree to which the evidence contradicts a proposition. It is called dubiety.

If p data sources are available, probability masses $m_i(B_j)$ have to be defined for each data source i with $1 \leqslant i \leqslant p$ and for all classes $B_j \in 2^\Theta$. The Dempster–Shafer theory allows the combination of these probability masses from several data sources to compute a combined probability mass for each class $A \in 2^\Theta$:

$$m(A) = \frac{\sum_{B_1 \cap B_2 \cdots \cap B_p = A} \left(\prod_{1 \leqslant i \leqslant p} m_i(B_j) \right)}{1 - \sum_{B_1 \cap B_2 \cdots \cap B_p = \emptyset} \left(\prod_{1 \leqslant i \leqslant p} m_i(B_j) \right)} \tag{8}$$

The sum in the denominator of Eq. (8) is a measure of the *conflict* in the evidence. As soon as the combined probability masses $m(A)$ have been derived from the original ones, $\mathrm{Sup}(A)$, $\mathrm{Pls}(A)$, and $\mathrm{Sup}(\overline{A})$ can be computed. Finally, a decision rule must be defined in order to determine the accepted simple hypothesis $C \in \Theta$. There are several ways of defining such a decision rule: C can be chosen to be the simple hypothesis (1) of maximum support, (2) of maximum plausibility, or (3) of maximum support without overlapping of uncertainty intervals [18]. We use the rule of maximum support in our application.

2.4.2. Definition of the probability masses

The definition of the probability masses is the distinguishing feature of any application of the Dempster–

Shafer theory. There are different strategies on how they can be defined. In [17], non-zero probability masses derived from probabilities from a previous classification are assigned to the simple hypotheses, and the only compound class receiving a non-zero probability mass is Θ, which is used to model the imprecision of the initial classifications. In [18] it is argued that such a definition is appropriate in cases where the information provided by the different sources is mainly redundant. The authors propose two strategies for assigning the probability masses in cases where the information from the different sources is complementary, which is particularly useful if two classes C_i and C_j cannot be distinguished by a certain sensor. In this case, a non-null probability mass should be assigned to $C_i \cup C_j$. Thus, $m(C_i \cup C_j) \neq 0$. C_i and C_j can either be assigned a zero probability mass, thus $m(C_i) = m(C_j) = 0$, or the same mass as $m(C_i \cup C_j)$, thus $m(C_i) = m(C_j) = m(C_i \cup C_j) \neq 0$. The first assignment assumes total ignorance about the membership of a pixel to either C_i or C_j. In [18], the second strategy is preferred for practical reasons: the paper deals with unsupervised classification, and applying the first strategy resulted in too many clusters in feature space.

We prefer the first strategy. Unlike the authors of [18], we do know the thematic classes we want to distinguish from the beginning, and their number is small. As we shall see in the subsequent sections, the cues (sensors) we use can distinguish two complementary subsets $B_1 \in 2^\Theta$ and $B_2 \in 2^\Theta$, with $B_1 \cap B_2 = \emptyset$ and $B_1 \cup B_2 = \Theta$. Both B_1 and B_2 consist of simple classes $C_{gh} \in \Theta$ with $B_1 = \{C_{11}, \ldots, C_{m1}\}$ and $B_2 = \{C_{12}, \ldots, C_{k2}\}$, m and k being the number of simple classes in B_1 and B_2, respectively. Of course, B_1 and B_2 can be different for different cues; otherwise, the information provided by the sensors would not be complementary. If neither B_1 nor B_2 consists of one simple class only, our ignorance about the simple classes C_{gh} is complete, and we can see no reason why we should assign a non-zero probability mass to any of them. Non-zero probability masses are thus only assigned to B_1 and B_2. Further, as B_1 and B_2 are mutually exclusive, $m(B_2) = 1 - m(B_1)$ holds true. As a consequence, the problem of defining probability masses is reduced to defining a probability mass function $P_i(x)$ for B_1 for each sensor i with $m_i(B_1) = P_i(x)$, where x is the output of sensor i. $P_i(x)$ can be interpreted as the probability of a certain pixel or feature to belong to class B_1 given that the output of sensor i equals x. It can also be interpreted as the result of an initial classification using only sensor i, and distinguishing only classes B_1 and B_2. In the following sections, we shall describe how the probability mass functions $P_i(x)$ are defined.

Table 1
The probability masses for the initial classification

Class	ΔH	R	D	ΔH_{FL}	NDVI	Combined probability mass
B	0	0	0	0	0	$\frac{P_{\Delta H} \cdot (1 - P_R) \cdot (1 - P_D) \cdot (1 - P_{\mathrm{FL}}) \cdot (1 - P_N)}{1 - C}$
T	0	P_R	P_D	P_{FL}	0	$\frac{P_{\Delta H} \cdot P_R \cdot P_D \cdot P_{\mathrm{FL}} \cdot P_N}{1 - C}$
G	0	0	0	0	0	$\frac{P_N \cdot (1 - P_{\Delta H}) \cdot (1 - P_R) \cdot (1 - P_D) \cdot (1 - P_{\mathrm{FL}})}{1 - C}$
S	0	0	0	0	0	$\frac{(1 - P_{\Delta H}) \cdot (1 - P_R) \cdot (1 - P_D) \cdot (1 - P_{\mathrm{FL}}) \cdot (1 - P_N)}{1 - C}$
$B \cup T$	$P_{\Delta H}$	0	0	0	0	0
$B \cup G$	0	0	0	0	0	0
$B \cup S$	0	0	0	0	$1 - P_N$	0
$T \cup G$	0	0	0	0	P_N	0
$T \cup S$	0	0	0	0	0	0
$G \cup S$	$1 - P_{\Delta H}$	0	0	0	0	0
$B \cup T \cup G$	0	0	0	0	0	0
$B \cup T \cup S$	0	0	0	0	0	0
$B \cup G \cup S$	0	$1 - P_R$	$1 - P_D$	$1 - P_{\mathrm{FL}}$	0	0
$T \cup G \cup S$	0	0	0	0	0	0
$B \cup T \cup G \cup S$	0	0	0	0	0	0

Classes: buildings (B), trees (T), grassland (G), and bare soil (S). ΔH: initial probability masses for the height differences in the normalised $\mathrm{DSM_L}$. R: initial probability masses for the strength of surface roughness. D: initial probability masses for the directedness of surface roughness. ΔH_{FL}: initial probability masses for the height differences between first and last pulses. NDVI: initial probability masses for the NDVI. The conflict C is the sum in the denominator of Eq. (8).

$$C = P_R \cdot (1 - P_{\Delta H} \cdot P_D) + P_D \cdot (1 - P_R) + P_{\Delta H} \cdot P_R \cdot P_D \cdot (1 - P_N) + P_{\Delta H} \cdot P_N \cdot (1 - P_R) \cdot (1 - P_D) + P_{\Delta H} \cdot P_{\mathrm{FL}} \cdot (1 - P_R) \cdot (1 - P_D) \cdot (1 - P_N) + P_{\mathrm{FL}} \cdot P_N \cdot (1 - P_{\Delta H}) \cdot (1 - P_R) \cdot (1 - P_D) + P_{\mathrm{FL}} \cdot (1 - P_{\Delta H}) \cdot (1 - P_R) \cdot (1 - P_D) \cdot (1 - P_N) + P_{\Delta H} \cdot P_R \cdot P_D \cdot P_N \cdot (1 - P_{\mathrm{FL}}).$$

2.4.3. Initial land cover classification

In this process, we want to achieve a per-pixel classification of the input data into one of four classes: Buildings (B), trees (T), grassland (G), and bare soil (S). Five cues derived from the original input data are used for this purpose. None of the five cues could be used individually to distinguish sharply between the four classes B, T, G, and S. They are considered to be five data sources in the Dempster–Shafer fusion process. Table 1 shows our definition of the initial probability masses according to the guidelines developed in Section 2.4.2 and the way they propagate to the final probability masses:

- The height differences ΔH between the $\mathrm{DSM_L}$ and the DTM can be used to distinguish elevated objects from the ground. In our classification scheme, the trees and the buildings are elevated objects, whereas the other classes represent the ground. We assign a probability mass $P_{\Delta H} = P_{\Delta H}(\Delta H)$ to the combined class $B \cup T$, and $(1 - P_{\Delta H})$ to $G \cup S$. $P_{\Delta H}$ is chosen to be an increasing function of ΔH.
- The strength R of surface roughness grows with the second derivatives of the DSM (cf. Eqs. (2) and (5)), thus with changes of the surface normal vectors. Large variations of the surface normal vectors are typical for trees, the only object class in our classification scheme not having a smooth surface, so that we use R to distinguish trees from other object classes. We assign a probability mass $P_R = P_R(R)$ to class T, and $(1 - P_R)$ to $B \cup G \cup S$. By that assignment, we neglect the fact that R can also be very large at the building boundaries and at the step edges of the terrain. P_R is chosen to be an increasing function of R.
- The directedness D of surface roughness depends on the local variations of the directions of the surface normal vectors. With trees, surface normal vectors usually do not change in an anisotropic way. That is why D is also an indicator for trees, but only if R is above a certain threshold (because otherwise, D might be dominated by the noise in a planar area). We assign a probability mass $P_D = P_D(D, R)$ to class T, and $(1 - P_D)$ to $B \cup G \cup S$. P_D is chosen to be an increasing function of D if R is greater than a certain threshold.
- The height differences ΔH_{FL} between the first and the last pulse DSMs are used in a similar way as the strength R of surface roughness. We neglect the large values of ΔH_{FL} at building boundaries and at power lines. We assign a probability mass $P_{\mathrm{FL}} = P_{\mathrm{FL}} \times (\Delta H_{\mathrm{FL}})$ to class T, and $(1 - P_{\mathrm{FL}})$ to $B \cup G \cup S$. P_{FL} is chosen to be an increasing function of ΔH_{FL}.
- The NDVI is an indicator for vegetation, thus for classes T and G. A probability mass $P_N = P_N\,(NDVI)$ is assigned to the combined class $T \cup G$, and $(1 - P_{\mathrm{FL}})$ to $B \cup S$. P_N is chosen to be an increasing function of the NDVI.

For modelling the probability masses $P_i(x)$ for these five cues, we use a heuristic approach that shares some similarities to the concept of membership functions in fuzzy logic [12]. We have described previously why each of the cues separates two subsets of Θ in our classification scheme. For input parameters x smaller than a threshold x_1, we assume the assignment of a pixel to class B_1 (cf. Section 2.4.2) to be very unlikely, which is modelled by a small probability mass P_1. For input parameters above a second threshold x_2 with $x_1 < x_2$, we assume the assignment of a pixel to class B_1 to be almost certain, which is modelled by a rather large probability mass P_2, with $0 \leqslant P_1 < P_2 \leqslant 1$. For instance, if ΔH is smaller than 1.5 m, it is very unlikely that there is a building or a tree; if it is greater than 3 m, it is very unlikely that there is grassland or bare soil. Between x_1 and x_2, the probability mass function should be defined in a way that there are no step edges (which would correspond to applying "hard" thresholds), but rather a smooth transition between the two probability levels P_1 and P_2. This could be a straight line, but we use a cubic parabola with horizontal tangents (thus being differentiable) at $(x = x_1)$ and at $(x = x_2)$ (Fig. 2). Thus, the probability masses $P_{\Delta H}$, P_R, P_{FL}, and P_N are computed according to Eq. (9):

$$P_i(x) = \begin{cases} P_1 & \forall x | x \leqslant x_1 \\ P_1 + (P_2 - P_1) \cdot \left[3 \cdot \left(\frac{x - x_1}{x_2 - x_1} \right)^2 - 2 \cdot \left(\frac{x - x_1}{x_2 - x_1} \right)^3 \right] & \\ \quad \forall x | (x > x_1) \wedge (x < x_2) & \\ P_2 & \forall x | x \geqslant x_2 \end{cases} \tag{9}$$

A slightly different definition has to be used for the probability mass function P_D because P_D is only significant if R is significant also. Thus, if R is below a threshold $R_{\min}$, P_D cannot contribute to the classification, which is modelled by assigning a probability mass of 0.5 to both T and $B \cup G \cup S$. Otherwise, the probability mass is also modelled by Eq. (9). We choose $R_{\min} = 5 \cdot \mathrm{median}(R)$. By the latter selection, the threshold is made adaptive to the average surface roughness of a scene (cf. Section 2.2).

We use $P_1 = 5\%$ and $P_2 = 95\%$. Note that $P_1 \neq 0\%$ and $P_2 \neq 100\%$. This is the main difference between our functional model and the membership functions for fuzzy logic described in [12]. We never assume the information from a sensor to be 100% certain. Conflicts

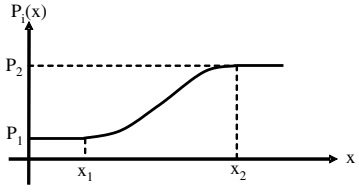

Fig. 2. The probability mass function.

Table 2
The values for (x_1, x_2) for $P_{\Delta H}$, P_R, P_D, P_{FL}, and P_N

	ΔH (m)	R	D	ΔH_{FL} (m)	NDVI (%)
x_1	1.5	$2 \cdot \mathrm{median}(R)$	0.1	1.5	30
x_2	3.0	$15 \cdot \mathrm{median}(R)$	0.9	3.0	65

in sensor information are expected both because the number of classes in our classification scheme is much smaller than the number of object classes that are actually observable in the data and the sensor data exhibit random variation. Actually, unlike other classification techniques, the Dempster–Shafer theory gives a direct measure of that conflict. If two sensors contradict each other and if the information of both sensors are considered to be 100% certain, that conflict cannot be resolved: the conflict would be equal to 1.0, and Eq. (8) could not be evaluated. We do not consider this to be a restriction of the Dempster–Shafer theory, but rather consider it to be an advantage that such a situation can be clearly detected.

The values we use for (x_1, x_2) are listed in Table 2. As with the value for P_D, they were determined empirically, but are assumed to be generally applicable. The values (x_1, x_2) for P_R are linked to the median of R and thus adaptive to the average slope variations in the data. The combined probabilities are evaluated for each pixel independently, and the pixel is assigned to the class of maximum support. In comparison to the rule-based algorithm described in [11] which is used for hierarchical DTM generation, there are several improvements.

- All cues are evaluated simultaneously.
- No sharp thresholds are required, and the probability mass function (Eq. (9)) has a smooth transition between the two levels P_1 and P_2.
- We do not eliminate parts of the data during the process as not belonging to the class "building", but we achieve an overall classification of land cover with respect to our four classes.

2.4.4. Final classification of building regions

After the initial classification, we obtain a binary building mask of all pixels classified as "building". As all pixels were classified independently from each other, only a small local neighbourhood contributed to their classification (via R and D), which causes classification errors:

Table 3
The probability masses for the final classification

Class	ΔH_a	H	P	NDVI_a	Combined probability mass
B	0	0	0	0	$\frac{P_{\Delta Ha} \cdot P_H \cdot (1-P_P) \cdot (1-P_{Na})}{1-C}$
T	0	$1-P_H$	P_P	0	$\frac{P_{\Delta Ha} \cdot P_P \cdot P_{Na} \cdot (1-P_H)}{1-C}$
G	0	0	0	0	$\frac{P_H \cdot P_{Na} \cdot (1-P_{\Delta Ha}) \cdot (1-P_P)}{1-C}$
S	0	0	0	0	$\frac{P_H \cdot (1-P_{\Delta H_a}) \cdot (1-P_P) \cdot (1-P_{Na})}{1-C}$
$B \cup T$	$P_{\Delta Ha}$	0	0	0	0
$B \cup G$	0	0	0	0	0
$B \cup S$	0	0	0	$1-P_{Na}$	0
$T \cup G$	0	0	0	P_{Na}	0
$T \cup S$	0	0	0	0	0
$G \cup S$	$1-P_{\Delta Ha}$	0	0	0	0
$B \cup T \cup G$	0	0	0	0	0
$B \cup T \cup S$	0	0	0	0	0
$B \cup G \cup S$	0	P_H	$1-P_P$	0	0
$T \cup G \cup S$	0	0	0	0	0
$BT \cup G \cup S$	0	0	0	0	0

Classes: buildings (B), trees (T), grassland (G), and bare soil (S). ΔH_a: initial probability masses for the average height of the building. H: initial probability masses for the percentage of "homogeneous" pixels. P: initial probability masses for the percentage of "point" pixels. NDVI_a: initial probability masses for the average NDVI. The conflict C is the sum in the denominator of Eq. (8).

$$C = (1-P_H) \cdot (1-P_{\Delta H_a} \cdot P_P) + P_P \cdot P_H + P_{\Delta Ha} \cdot P_P \cdot (1-P_H) \cdot (1-P_{Na}) + P_{\Delta H_a} \cdot P_H \cdot P_{Na} \cdot (1-P_P).$$

- There are singular "building" pixels inside larger areas of other classes, or "tree" pixels inside building roofs.
- R and D give relatively large values for small detached houses if the resolution of the LIDAR data is not better than or equal to about 0.5 m. This might result in classification errors that cannot be corrected by a local analysis.

To overcome these problems, isolated building pixels are eliminated by binary morphological opening, and a building label image is created by a connected component analysis. After that, the Dempster–Shafer theory is applied for a final classification of the original building regions thus detected, this time using four cues representing average values for each building region (Table 3):

- The average height differences ΔH_a between the $\mathrm{DSM_L}$ and the DTM are used in a similar way as ΔH in the original classification by assigning a probability mass $P_{\Delta H_a} = P_{\Delta H_a}(\Delta H_a)$ to $B \cup T$, and $(1 - P_{\Delta H_a})$ to $G \cup S$.
- The percentage H of pixels classified as "homogeneous" in polymorphic feature extraction (cf. Section 2.2) is an indicator for objects consisting of planar surface patches. Thus, we assign a probability mass $P_H = P_H(H)$ to class $B \cup G \cup S$, and $(1 - P_H)$ to class T.
- The percentage P of pixels classified as "point-like" in polymorphic feature extraction is an indicator for trees. We assign a probability mass $P_P = P_P(P)$ to class T, and $(1 - P_P)$ to $B \cup G \cup S$.
- The average NDVI_a is used in a similar way as in the original classification by assigning a probability mass $P_{Na} = P_{Na}(NDVI_a)$ to $T \cup G$, and $(1 - P_{Na})$ to $B \cup S$.

The height differences between first and last pulses are no longer used. For the probability masses $P_{\Delta H_a}$, P_H, P_P, and P_{Na} we again use the function described by Eq. (9), with $P_1 = 5\%$ and $P_2 = 95\%$. The values for (x_1, x_2) are presented in Table 4. These values are designed to avoid eliminating buildings erroneously, but to accept a larger false-alarm rate. The combined probability masses are evaluated for each initial building region, and if such a region is assigned to a class other than "building", it is eliminated. Finally, the regions classified as buildings are grown by a few pixels by morphological filtering to correct for building boundaries being erroneously classified as trees, caused by the large values for R (cf. Section 2.4.3). The advantage of this method compared to the one described in [11] is that it considers all cues simultaneously and avoids sharp thresholds.

Table 4
The values for (x_1, x_2) for $P_{\Delta H_a}$, P_H, P_P, and P_{Na}

	ΔH_a (m)	H (%)	P (%)	NDVI_a (%)
x_1	1.5	0	30	30
x_2	3.0	60	75	65

3. Experiments

3.1. Description of the data set

The test data set was captured over Fairfield (NSW) using an Optech ALTM 3025 laser scanner. It covers an area of 2×2 km^2. Both first and last pulses and intensities were recorded with an average point distance of about 1.2 m. We derived DSM grids at a resolution of 1 m from these data. A true colour digital orthophoto with a resolution of 0.15 m was also available for the area. The orthophoto had been created using a DTM, so that the roofs and the tree-tops were displaced with respect to the LIDAR data. Thus, data alignment was not perfect. Unfortunately, the digital orthophoto did not contain an infrared band. We circumvented this problem by resampling both the digital orthophoto and the (infrared) LIDAR intensity data to a resolution of 1 m and by computing a "pseudo-NDVI-image" from the LIDAR intensities and the red band of the digital orthophoto. Apart from problems with data alignment caused by the displaced tree canopies in the orthophoto, there were also problems with shadows in the orthophoto, so the pseudo-NDVI image did not provide as much information as expected.

A reference data set was created by digitising building polygons interactively in the digital orthophoto. We chose to digitize all structures recognisable as buildings independent of their size. The reference data include garden sheds, garages, etc., that are sometimes smaller than 10 m^2 in area. Such small structures cannot be expected to be detected in the LIDAR data, given their resolution. Neighbouring buildings that were joined but are obviously separate entities were digitized as separate polygons. Thus, altogether 2385 polygons were digitized. As the aerial image used for producing the orthophoto and the LIDAR data were captured at different epochs, there were contradictions in the two data sets as some buildings were either constructed or demolished between them. Altogether 48 polygons were detected in only one data set. Of these, all except one (an industrial building) were larger detached houses. Smaller entities could not be checked in the LIDAR data. The DSM and the pseudo-NDVI-image as well as a label image created from the reference data set are shown in Fig. 3.

Fig. 3. The Fairfield data set. Left: the DSM from last pulse data (black: low areas, white: high areas). Centre: The pseudo-NDVI image. Right: a label image derived from the reference data set. Total area: 2000×2000 m^2.

3.2. Method of evaluation

In the evaluation process, we compared two data sets: the "automatic data set" consisting of building regions detected automatically, and the reference data set. The comparison of two spatial data sets is not a straightforward task if their topologies are different [22]. This is the case here because the automatic building detection process cannot separate buildings that are actually joined in object space or that are so close to each other that there are no LIDAR points between them. We have to expect buildings to be merged in the detection process. Thus, a comparison of the boundary polygons proposed in [22] is not appropriate. Alternatively, we compare two label images: the label image that is the output of automatic building detection, hence called the "automatic label image", and a label image created from the polygons of the reference data set with the same spatial resolution as the automatic label image, hence called the "reference label image". Before the actual evaluation process, the areas covered by the polygons detected to be available in one data set only are erased in both of these label images. For an evaluation of automatic feature extraction using a reference data set, two numbers of interest are the *completeness* and the *correctness* of the results [23]:

$$\text{Completeness} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

$$\text{Correctness} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

In Eqs. (10) and (11), TP denotes the number true positives, i.e., the number of entities found to be available in both data sets. FN is the number of false negatives, i.e., the number of entities in the reference data set that were not detected automatically, and FP is the number of false positives, i.e., the number of entities that were detected, but do not correspond to an entity in the reference data set. We are interested in determining completeness and correctness for two types of entities. First, we want to derive them on a per-pixel level. In this context, TP is the number of pixels classified as "building" in both label images, FN is the number of building pixels in the reference label image not classified as "building" in the automatic label image, and FP is the number of building pixels in the automatic label image not classified as "building" in the reference label image. The numbers thus derived for completeness and correctness give a balanced estimate of the area that has been correctly classified as "building". Second, we are interested in numbers on a per-building level, showing how many buildings could be detected and how many of the buildings detected automatically did actually correspond to buildings. In this case, the TP, FN, and FP cannot be determined easily because of the problem of multiple and partial overlaps of building regions in the automatic and in the reference data sets.

We denote the sets of regions from the automatic and the reference label images by L_a and L_r, respectively. For each co-occurrence of two labels $l_a \in L_a$ and $l_r \in L_r$, we compute the overlap ratios $p_{ar} = n_{a\cap r}/n_a$ and $p_{ra} = n_{a\cap r}/n_r$, where $n_{a\cap r}$ is the number of common pixels assigned to the region l_a in the automatic label image and to l_r in the reference label image, n_a is the total number of pixels assigned to the region l_a in the automatic label image, and n_r is the total number of pixels assigned to l_r in the reference label image. Obviously, if the two data sets were characterised by the same topology, both p_{ar} and p_{ra} would be close to 100% for all building regions, and there would be exactly one region $l_a \in L_a$ corresponding to each $l_r \in L_r$ and vice versa. As this is not the case, we have to evaluate the overlap percentages p_{ar} and p_{ra} further, to match corresponding regions in the two data sets.

Initially, all tuples (l_a, l_r) with $n_{a\cap r} > 0$ are considered to be possible correspondences. Our analysis starts by eliminating spurious correspondences. We define a function overlap(l_i, l_j) classifying the overlap between regions $l_i \in L_i$ and $l_j \in L_j$ with $i \in \{r, a\}, j \in \{r, a\}$, and $i \neq j$:

$$\text{overlap}(l_i, l_j) = \begin{cases} \text{strong} & \forall i,j | p_{ij} > 80\% \\ \text{partial} & \forall i,j | 80\% \leqslant p_{ij} < 50\% \\ \text{weak} & \forall i,j | 50\% \leqslant p_{ij} < 10\% \\ \text{none} & \forall i,j | p_{ij} \leqslant 10\% \end{cases} \quad (12)$$

The function defined in Eq. (12) is not necessarily symmetric, thus we cannot expect $\text{overlap}(l_a, l_r)$ to be identical to $\text{overlap}(l_r, l_a)$. We consider a correspondence between two regions $l_a \in L_a$ and $l_r \in L_r$ to be spurious if both $\text{overlap}(l_a, l_r)$ and $\text{overlap}(l_r, l_a)$ are either *weak* or *none*. These correspondences are no longer considered. For each region $l_a \in L_a$ we obtain a subset $L_{ar} \subset L_r$ containing all regions from L_r that correspond to l_a:

$$L_{ar} = \{l_r \in L_r | [\text{overlap}(l_r, l_a) \in \{\text{strong}, \text{partial}\}] \vee [\text{overlap}(l_a, l_r) \in \{\text{strong}, \text{partial}\}]\} \quad (13)$$

In the same way, for each region $l_r \in L_r$ we obtain the subset $L_{ra} \subset L_a$ of corresponding regions from L_a:

$$L_{ra} = \{l_a \in L_a | [\text{overlap}(l_r, l_a) \in \{\text{strong}, \text{partial}\}] \vee [\text{overlap}(l_a, l_r) \in \{\text{strong}, \text{partial}\}]\} \quad (14)$$

L_{ra} can be interpreted as the set of regions of the automatic data set into which a region l_r of the reference data set is split. L_{ar} is the set of regions of the reference data set which are merged into a region l_a of the automatic data set. Having found corresponding regions, the overall coverage d_i for each region $l_i \in L_i$ can be computed with $i \in \{r, a\}, j \in \{r, a\}$, $i \neq j$, and $n_{a \cap r} = n_{r \cap a}$:

$$d_i = \frac{\sum_{l_j \in L_{ij}} n_{i \cap j}}{n_i} \quad (15)$$

Thus, d_i is the ratio between the sum of the number of all pixels overlapping with one of the corresponding regions of the other data set and the total number of pixels of region l_i. For a region $l_r \in L_r$, d_r is the percentage of the area of l_r that is substantially covered by regions detected automatically. We consider a region l_r to be *completely detected* if $d_r > 80\%$, *partly detected* if $80\% \leqslant d_r < 50\%$, *hardly detected* if $50\% \leqslant d_r < 10\%$, and *not detected* if $d_r \leqslant 10\%$. For a region $l_a \in L_a$, d_a is the percentage of the area of l_a that actually corresponds to regions of the reference data set. We consider a region l_a to be *completely correct* if $d_a > 80\%$, *partly correct* if $80\% \leqslant d_a < 50\%$, *hardly correct* if $50\% \leqslant d_a < 10\%$, and *not correct* if $d_a \leqslant 10\%$. This gives us the tools for computing the numbers of TP, FN, and FP in Eqs. (10) and (11):

- In Eq. (10), TP is the number of building regions in the reference data set that are either *completely detected* or *partly detected*.
- In Eq. (11), TP is the number of building regions in the automatic data set that are either *completely correct* or *partly correct*.
- FN is the number of building regions in the reference data set that are neither *completely detected* nor *partly detected*.
- FP is the number of building regions in the automatic data set that are neither *completely correct* nor *partly correct*.

Note that the different definitions of TP for computing completeness and correctness is a consequence of the fact that we consider regions that are either split or merged to be correct if the overall coverage is sufficient. We assume a region with coverage larger than 80% to be completely detected/correct because we have to consider errors at the building boundaries, which might comprise a considerable percentage of smaller buildings, and we believe that such errors can be corrected in a later stage of processing if the delineations of the roof planes are searched [3].

3.3. Results

Fig. 4 shows the normalised DSM created by three iterations of morphological opening, using structural elements of 150, 75, and 25 m. In the second iteration, buildings larger than 225 m^2, containing at least 45% of homogeneous and less than 20% of point-like pixels are detected. They are eliminated in the third iteration of DTM generation, so that they are preserved in the nDSM in Fig. 4.

Fig. 5 depicts the probability masses for the initial Dempster–Shafer classification. Note that $P_{\Delta H}$ is large both for buildings and trees, whereas P_R, P_D and P_{FL} give large values for trees. P_R also highlights the building boundaries. The areas in a medium grey in P_D are those where P_D was set to 0.5 because the texture was not considered to be significant. In these areas, P_D did not contribute to the classification. The dominant linear structures in P_{FL} are powerlines, which are not considered by our approach. P_N distinguishes bare soil and,

Fig. 4. The normalised DSM used in Dempster–Shafer classification (black: low areas, white: high areas).

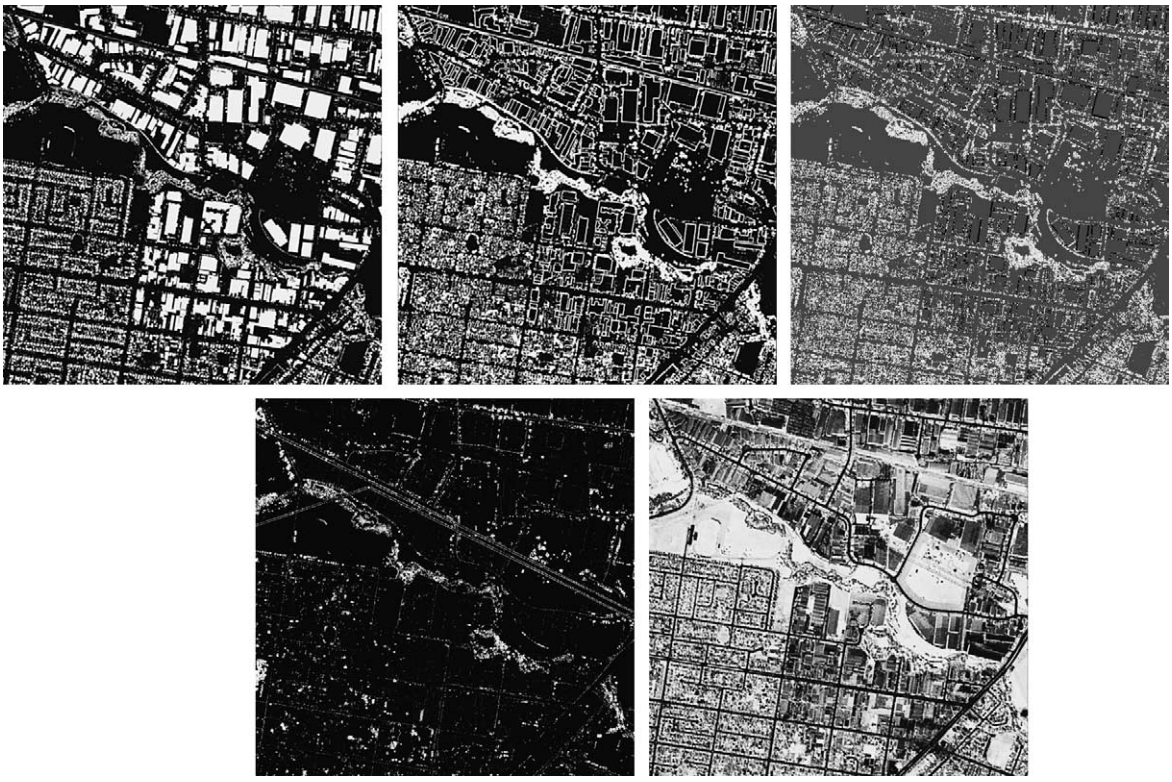

Fig. 5. The probability masses for the initial Dempster–Shafer classification. Upper row: $P_{\Delta H}$ (left), P_R (centre), and P_D (right); second row: P_{FL} (left) and P_N (right).

Fig. 6. The results of the initial Dempster–Shafer classification. White: grass-land (G), light grey: bare soil (S), dark grey: trees (T), black: buildings.

Fig. 7. The final building label image after the second Dempster–Shafer classification and after growing the building regions by morphological closing.

less clearly, buildings from vegetation. Note that the industrial buildings have a relatively large pseudo-NDVI. The results of Dempster–Shafer classification are presented in Fig. 6.

After morphological opening of the binary image of the building pixels from Fig. 6 and after eliminating building candidate regions smaller than 10 m^2, the second Dempster–Shafer classification is carried out for

altogether 2291 building candidate regions detected in the data. 344 of these regions are found to belong to a class other than "building", so that we finally obtain 1947 building regions. Fig. 7 shows the final building label image after growing the building regions by morphological closing to compensate for the incorrect classification of the building boundaries. The computation time for achieving the result in Fig. 7 was about 5 min on a Pentium 4 PC (2.66 GHz, 512 MB RAM).

3.4. Evaluation of the results and discussion

By evaluating the results of the initial classification using the Dempster–Shafer technique in Fig. 6, it is clear that the class "bare soil" mainly corresponds to streets and parking lots. Most of the trees are situated in the valley of the river crossing the scene diagonally, along the streets in the residential area in the south-eastern part of the scene, and in the backyards of the houses. Step edges at the building boundaries are often classified as trees. Given the resolution of the LIDAR data, it was not easy to separate trees from buildings in the residential areas, which is also the reason for the rugged appearance of the building boundaries in these regions. A few of those residential buildings were erroneously classified as trees, especially if the roof consisted of many small faces. Further problems occurred with bridges because morphological filtering results in a DTM corresponding to the terrain below the bridge. This results in large height differences ΔH that indicate the presence of a building or a tree, whereas the surface roughness of the street and the NDVI indicate an area not covered by trees or vegetation, so that the overall classification would assign such areas to the class "building". There are also isolated spots on top of the large industrial buildings that are not classified correctly, which is mostly caused by chimneys or other objects yielding a large local variation of the surface normal vectors. In the parking lots, there are errors where cars are parked. Of course, cars were not considered in our classification scheme. Some dominant power lines are still partly preserved in the data, classified as "bare soil" inside grassland areas, which is caused by the low reflectance of the power lines in the wavelength of the laser scanner. In general, shadows in the colour orthophoto where an error source, because no shadows appeared in the LIDAR intensity data, so that the "pseudo-NDVI-image" gave systematically wrong NDVI values in these areas. Finally, there are also some incorrectly classified isolated pixels inside larger areas of another class, because context was not considered in the first classification process.

The parameter settings for that second classification seem to be more critical and more dependent on the data than those for the first, because with increasing resolution of the LIDAR data, the percentage of "homogeneous" pixels in roof planes will also increase, whereas the percentage of "point" pixels will become smaller. This implies that the second classification stage will be more helpful for the discrimination of buildings and trees with LIDAR data of a higher resolution than it is in our test project. This claim is confirmed in [3].

Fig. 8. Evaluation of the results of building detection. Light grey: correct building pixels; medium grey: false positives; black: false negatives.

Fig. 8 shows the results of the evaluation of the automatic building extraction process. On a per-pixel basis, the *completeness* was 94%, thus 94% of the building pixels were actually detected, which is very satisfactory. The missed buildings (black areas in Fig. 8) were small residential buildings, some having roofs with high reflectance in the wavelength of the laser scanner (thus, a high pseudo-NDVI), others having roofs consisting of many small planar faces, or they are too small to be detected given the resolution of the LIDAR data. For a few large industrial buildings, some building parts could not be detected due to errors in DTM generation. However all large buildings except one (at the upper margin) were substantially detected. The *correctness* on a per-pixel basis was 85%, thus 85% of the pixels classified as building pixels do actually correspond to a building. This is not quite as good as the completeness. Fig. 8 shows that this number is affected by errors at the building boundaries. After enlarging the buildings using the approach described in Section 2.4.4, the buildings seem to be slightly too large. There are also a few larger false positives at bridges, at small terrain structures not covered by vegetation, in areas with overseas containers, and in trailer parks.

The results of the evaluation on a per-building basis are presented in Fig. 9. The upper diagram shows both completeness and correctness depending on the area

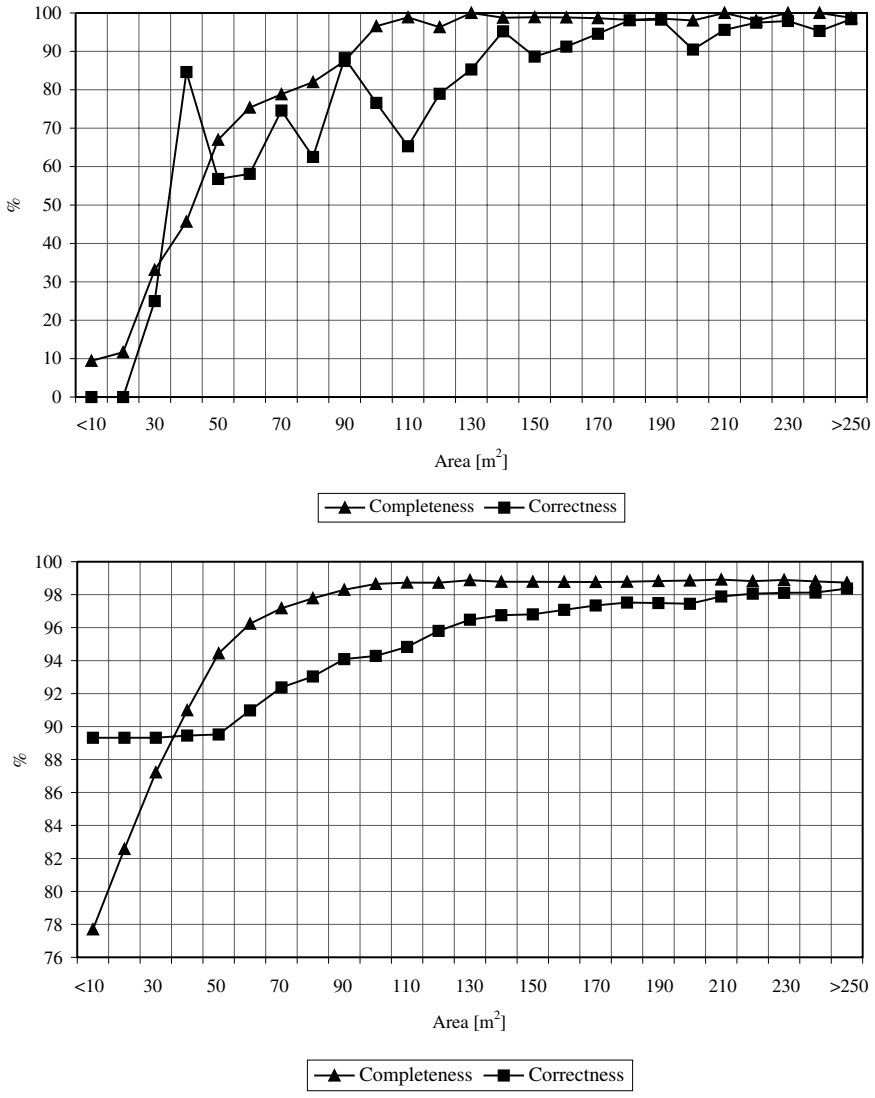

Fig. 9. Above: completeness and correctness of buildings depending on the building size (class width: 10 m^2). Below: cumulative completeness and correctness of buildings larger than the size shown in the abscissa.

covered by the buildings using a class width of 10 m^2, i.e., it shows completeness and correctness computed separately for all area intervals of the abscissa. It is obvious that the quality of the results depends on the building size: for buildings of a size larger than 90 m^2, completeness is greater than 90%, but it becomes less than 50% for buildings smaller than 40 m^2. A similar trend can be observed for the correctness, which is again not as good as completeness. For buildings larger than 130 m^2, correctness is greater than 90%. For buildings between 130 and 40 m^2, correctness oscillates between 90% and 50%, and it rapidly descends toward zero for detected building regions smaller than 40 m^2. The lower diagram in Fig. 9 shows the cumulative completeness and correctness for buildings covering an area larger than the size shown in the abscissa. It shows that 95% of all buildings larger than 50 m^2 and 90% of all buildings larger than 30 m^2 could be detected, whereas 96% of the detected building regions larger than 120 m^2 and 89% of all detected building regions were correct. The number of small buildings in the reference data set was relatively large: 570 buildings or 24% covered an area smaller than 50 m^2, 296 or 12% were smaller than 30 m^2. On the other hand, only 30 buildings or 1.5% of the detected regions were smaller than 50 m^2. Based on Figs. 8 and 9, it is clear that the main buildings (larger than 90 m^2) were detected reliably by our method. The majority of the buildings between 50 and 90 m^2 could also be detected, whereas buildings smaller than 30 m^2 (mostly garden sheds or garages) could not usually be detected.

A comparison of these results with those presented in literature is difficult for two reasons: first, the test data are not identical, so that any comparison is somewhat uncertain, and second, the methodologies used for evaluation are not standardised and thus are often different. For instance, most authors do not give detection

rates depending on the building size, and others only give per-pixel evaluations. In [12], buildings were extracted from LIDAR data of 1 m resolution in rural and urban test areas, with detection rates of 92.6% and 95.8%, respectively. No information is given about the sizes of the buildings in the rural area (that could be compared to ours), and we do not know how the reference data were collected. The authors also state that the classification accuracy decreases with the building size. In [24], building change detection was carried out using LIDAR data sampled at 0.6 m. The authors state that 90% of all building pixels in a reference map were correctly detected, whereas 80.3% of the detected building pixels were also building pixels in the reference map. These numbers correspond to our completeness and correctness numbers, and are of a similar size. The authors also give completeness and correctness numbers separately for buildings larger than 200 m^2 and smaller than that threshold. Their numbers confirm the trend that can be seen in Fig. 9: completeness and correctness are 91.1% and 84.1%, respectively, for buildings larger than 200 m^2. For buildings larger than 200 m^2, completeness and correctness are given by 42.1% and 34.9%, respectively. A minimum percentage of overlap of 70% is required for a building to be correctly identified [24]. We believe that the results in our tests demonstrate similar if not better results than these tests, and hence justify the approach taken for automatic building extraction in this research.

4. Conclusions and future work

We have presented a method for building detection from LIDAR data and multi-spectral images, and we have shown its applicability in a test site of heterogeneous building shapes. The method is based on a hierarchical technique for DTM generation and on the application of the Dempster–Shafer theory for classification. The results achieved were very satisfactory. 95% of the buildings larger than 50 m^2 can be detected, whereas about 89% of the detected buildings are correct. The detection rates decrease considerably with the building size: building structures smaller than 30 m^2 could generally not be detected. As a general trend, there were more false positives than false negatives. The quality of our results is comparable to those achieved by other research groups. Although this comparison is not conclusive because no common data set was used and because the evaluation methods are not always comparable, it shows that the Dempster–Shafer theory is well-suited for building detection.

Future work will concentrate on the influence of the parameters of the probability mass function used in this study on the results. We are also interested in the relative contribution of the individual cues to the classification results, especially in improving the classification results by the multi-spectral data. Another topic of future research is the replacement of the heuristic model for the probability mass functions by an empirical one that could be derived from the original classification results in an iterative procedure. Moreover, an investigation of the classification accuracy in dependence of the LIDAR resolution would be interesting. We expect the second Dempster–Shafer classification stage to give better results with data of a higher resolution. We also expect the classification to be better if real near-infrared data are used, which is still to be investigated. Finally, we do not want the results of building detection as presented in this paper to be the end of this research. These results are a prerequisite to geometrical reconstruction of buildings from roof planes, which we should be able to detect from building regions by fusing aerial imagery and LIDAR data. We also want to use the results of the first Dempster–Shafer classification and the final results of building extraction to improve the quality of the DTM by eliminating points on the building roofs before applying robust linear prediction as described in [10].

Acknowledgements

This work was supported by the Australian Research Council (ARC) under Discovery Project DP0344678 and Linkage Project LP0230563. The Fairfield data set was provided by AAM Geoscan, Kangaroo Point, QLD 4169, Australia (www.aamgeoscan.com.au).

References

[1] K. Kraus, Principles of airborne laser scanning, J. Swedish Soc. Ph & RS 1 (2002) 53–56.
[2] N. Haala, C. Brenner, Extraction of buildings and trees in urban environments, ISPRS J. Ph & RS 54 (1999) 130–137.
[3] F. Rottensteiner, C. Briese, A new method for building extraction in urban areas from high-resolution LIDAR data, IAPSIS XXXIV 3A (2002) 295–301.
[4] G. Vosselman, S. Dijkman, 3D building model reconstruction from point clouds and ground plans, IAPSIS XXXIV 3W4 (2001) 37–43.
[5] G. Vosselman, On the estimation of planimetric offsets in laser altimetry data, IAPSIS XXXIV 3A (2002) 375–380.
[6] T. Schenk, B. Csatho, Fusion of LIDAR data and aerial imagery for a more complete surface description, IAPSIS XXXIV 3A (2002) 310–317.
[7] L. Klein, Sensor and Data Fusion, Concepts and Applications, SPIE Optical Engineering Press, 1999.
[8] H.-G. Maas, Fast determination of parametric house models from dense airborne laserscanner data, IAPRS XXXII 2W1 (1999) 1–6.
[9] U. Weidner, W. Förstner, Towards automatic building reconstruction from high resolution digital elevation models, ISPRS J. Ph & RS 50 (4) (1995) 38–49.

[10] C. Briese, N. Pfeifer, P. Dorninger, Applications of the robust interpolation for DTM determination, IAPSIS XXXIV 3A (2002) 55–61.
[11] F. Rottensteiner, J. Trinder, S. Clode, K. Kubik, Building detection using LIDAR data and multispectral images, in: Proc. APRS Conference on Digital Image Computing: Techniques and Applications, DICTA, Sydney, December 2003, vol. II, pp. 673–682.
[12] T. Voegtle, E. Steinle, On the quality of object classification and automated building modeling based on laserscanning data, IAPRSIS XXXIV 3W13 (2003) 149–155.
[13] A. Brunn, U. Weidner, Extracting buildings from digital surface models, IAPRS XXXII 3–4W2 (1997) 27–34.
[14] W. Förstner, A framework for low level feature extraction, in: J.O. Eklundh (Ed.), Computer Vision—ECCV '94, vol. II, 5th ICCV, Boston, MA, 1994, pp. 383–394.
[15] E. Steinle, T. Vögtle, Effects of different laser scanning modes on the results of building recognition and reconstruction, IAPRS XXXIII B3 (2000) 858–865.
[16] Y.H. Lu, J. Trinder, Data fusion applied to automatic building extraction in 3D reconstruction, in: Annual ASPRS Conference, Anchorage, Alaska, May 2003, pp. 114–122.
[17] T. Lee, J.A. Richards, P.H. Swain, Probabilistic and evidential approaches for multisource data analysis, IEEE Trans & RS GE 25 (3) (1987) 283–293.
[18] S. Le Hégarat-Mascle, I. Bloch, D. Vidal-Madjar, Application of Dempster–Shafer evidence theory to unsupervised classification in multisource remote sensing, IEEE Trans. Geosc. & RS 35 (4) (1997) 1018–1031.
[19] E. Binaghi, I. Gallo, P. Madella, A. Rampini, Approximate reasoning and multistrategy learning for multisource remote sensing data interpretation, in: C.H. Chen (Ed.), Information Processing for Remote Sensing, World Scientific Publishing, Singapore, 1999, pp. 397–429.
[20] B. Gorte, Supervised image classification, in: A. Stein, F. van der Meer, B. Gorte (Eds.), Spatial Statistics for Remote Sensing, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 153–163.
[21] R. Brügelmann, W. Förstner, Noise estimation for color edge extraction, in: W. Förstner, S. Ruwiedel (Eds.), Robust Computer Vision, Wichmann, Karlsruhe (Germany), 1992, pp. 90–107.
[22] L. Ragia, S. Winter, Contributions to a quality description of areal objects in spatial data sets, ISPRS J. Ph & RS 55 (3) (2000) 201–213.
[23] C. Heipke, H. Mayer, C. Wiedemann, O. Jamet, Evaluation of automatic road extraction, IAPRS XXXII (1997) 47–56.
[24] L. Matikainen, J. Hyyppä, H. Hyyppä, Automatic detection of buildings from laser scanner data for map updating, IAPSIS XXXIV 3W13 (2003) 218–224.