Int J Appl Earth Obs Geoinformation 70 (2018) 25–34

Contents lists available at ScienceDirect

Int J Appl Earth Obs Geoinformation

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

A patch-based method for the evaluation of dense image matching quality

Zhenchao Zhang[a,*], Markus Gerke[b], George Vosselman[a], Michael Ying Yang[a]

[a] Department of Earth Observation Science, Faculty ITC, University of Twente, Enschede, The Netherlands
[b] Institute of Geodesy and Photogrammetry, Technical University of Brunswick, Germany

ARTICLE INFO

Keywords:
Quality evaluation
Photogrammetry
Dense image matching
Laser scanning
Point cloud
Digital Surface Model

ABSTRACT

Airborne laser scanning and photogrammetry are two main techniques to obtain 3D data representing the object surface. Due to the high cost of laser scanning, we want to explore the potential of using point clouds derived by dense image matching (DIM), as effective alternatives to laser scanning data. We present a framework to evaluate point clouds from dense image matching and derived Digital Surface Models (DSM) based on automatically extracted sample patches. Dense matching errors and noise level are evaluated quantitatively at both the local level and whole block level. In order to demonstrate its usability, the proposed framework has been used for several example studies identifying the impact of various factors onto the DIM quality. One example study proves that the overall quality on smooth ground areas improves when oblique images are used in addition. This framework is then used to compare the dense matching quality on three different terrain types. In another application of the framework, a bias between the point cloud and the DSM generated from a photogrammetric workflow is identified. The framework is also used to reveal inhomogeneity in the distribution of the dense matching errors caused by overfitting the bundle network to ground control points.

1. Introduction

Airborne laser scanning (ALS) and airborne photogrammetry are the two main techniques to obtain 3D data representing the earth surface (Höhle and Höhle, 2009). The properties of laser scanning and photogrammetry have been widely compared before (Baltsavias, 1999; Leberl et al., 2010; Haala et al., 2010; Remondino et al., 2014; Cavegn et al., 2014; Yang and Chen, 2015; Tian et al., 2017). Compared to airborne laser scanning, image acquisition in airborne photogrammetry is mostly cheaper and more efficient in data acquisition flights (Hobi and Ginzler, 2012; Nurminen et al., 2013; Maltezos et al., 2016). In many countries photogrammetric image blocks are captured anyway for administrative and planning purposes with decreasing time intervals, so the question is to what extent these data can be used to replace ALS data in various application domains such as Digital Elevation Model (DEM) acquisition (Ressl et al., 2016), forestry mapping (Mura et al., 2015), classification and object extraction (Tomljenovic et al., 2016; Dong et al., 2017), and 3D modeling (Xiong et al., 2015).

We want to explore the potential of using photogrammetric products as effective alternatives to laser scanning data. In order to judge this potential, it is necessary to evaluate the data quality of 3D products from dense image matching (DIM). Assessing the absolute accuracy of 3D data can be time-consuming and labor-intensive for two reasons. Firstly, the reference data must be verified as being more accurate than the compared data. Secondly, the sample size should be sufficiently large in order to arrive at sound conclusions. Previous work of evaluating the absolute accuracy of 3D data can be divided into two categories based on the reference data.

In some previous evaluation studies, the reference data was collected by Real Time Kinematic (RTK) GPS. However, the sample size was relatively small in this case. Jaud et al. (2016) evaluated point clouds generated from images obtained by Unmanned Aerial Vehicles (UAVs). Twenty-four ground targets were set in the study area which served as GCPs in the triangulation and as check points in the DIM evaluation. The coordinates of these targets were obtained by post-processed differential GPS. Hobi and Ginzler (2012) evaluated the quality of Digital Surface Models (DSMs) from stereo matching of WorldView-2 satellite images and ADS80 aerial images using 36 reference points obtained by sub-decimeter differential GPS. Nurminen et al. (2013) studied the accuracy of DSMs derived from ALS and DIM in the estimation of plot-level variables. The reference variables of the forest plots were obtained by field surveys.

In addition, the reference data may be obtained by laser scanning. The basic assumption is that the point clouds obtained by laser scanning are more accurate than point clouds from photogrammetry, at least concerning the height component. Mandlburger et al. (2017) calculated the deviation between DIM-DSM and Lidar-DSM at impervious surfaces and found a systematic deviation of 0.043 m and a dispersion of

* Corresponding author.
E-mail address: z.zhang-1@utwente.nl (Z. Zhang).

https://doi.org/10.1016/j.jag.2018.04.002
Received 29 January 2018; Received in revised form 3 April 2018; Accepted 4 April 2018
Available online 11 April 2018

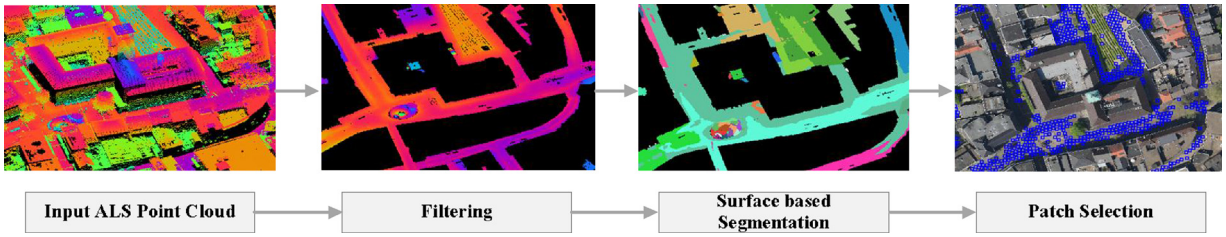


Fig. 1. Workflow for detecting candidate patches from ALS point cloud.

0.041 m. Tian et al. (2017) selected 184 inventory plots as the samples for DSM evaluation in a forest area. Two datasets from ALS were taken as reference data. Similar work taking laser scanning data as reference can also be found in (Poon et al., 2005; Gehrke et al., 2010; Moussa et al., 2013; Remondino et al., 2014; Nex et al., 2015; Jaud et al., 2016; Maltezos et al., 2016; Sofia et al., 2016; Ressl et al., 2016).

Some deficiencies of previous DIM evaluation work are summarized as follows: Firstly, some studies evaluated the point cloud derived from Semi-Global Matching (SGM) by making comparisons with ALS data or terrestrial laser scanning data on a planar sports field, complex castle or building façade (e.g. in Rothermel et al., 2012; Haala and Rothermel, 2012; Cavegn et al., 2014; Remondino et al., 2017). However, the small sample size or local area cannot properly represent the error distribution in the whole block. Secondly, when calculating quality measures, point-to-point distance (Kraus et al., 2006) and point-to-plane distance (Rothermel and Haala, 2011; Nex et al., 2015) were widely used as the measures to represent the accuracy. However, these measures are sensitive to blunders and random noise within the dense matching point clouds. Thirdly, the quality measures were less reliable or persuasive if calculated without consideration of the breaklines in natural scenes, such as bumpy terrain, edges of traffic islands or curbstones, and edges and ridges of roofs (e.g. in Ressl et al., 2016; Jaud et al., 2016).

In our previous work of evaluating point cloud from multi-view photogrammetry (Zhang et al., 2017), robust quality measures were calculated on roof segments. In this paper, a framework for evaluating point clouds and DSMs generated from a state-of-the-art dense matching algorithm is proposed. The contributions are as follows:

- The dense matching quality is evaluated robustly based on a very large number of planar patches of the same size extracted from planar ground surfaces in both the DIM point cloud and the ALS point cloud. Quantitative quality measures are proposed to represent the accuracy and precision at both the local patch level and the whole block level. After considering possible breaklines in natural scene and excluding patches with possible changes between the DIM data and reference data, the evaluation based on these planar patches reveals the distribution of DIM errors in the whole photogrammetric block for the first time. Compared to the previous point-to-point and point-to-plane comparisons, this framework computing the plane-to-plane distance is more robust to local blunders and artefacts.
- In order to test the usability of the proposed framework, several influencing factors related to the DIM quality are studied. One example is the additional use of oblique airborne imagery. This is not yet standard, but especially in urban applications it becomes more important (Toschi et al., 2017). Among other factors we evaluate how the additional use of oblique images influences the dense matching quality. We also compare the dense matching quality on different types of terrain. Meanwhile, suggestions are given on the photogrammetric quality control and dense matching parameter settings.

The paper is organized as follows: Section 2 presents the patch-based DIM evaluation framework. Section 3 gives details on the study area and experimental settings, while Section 4 focuses on experimental results. Section 5 discusses those results and Section 6 finally concludes the paper.

2. Methodology

In our evaluation framework, an ALS point cloud is taken as the reference data. The ALS data are assumed to be accurate with regards to the external reference and precise in consideration of random noise. The “patches” used as evaluation units are regular squares selected from the ALS data. Every patch is a sample for quality evaluation. Therefore, the densely selected patches on the ground can indicate the error distribution in the whole photogrammetric block. The proposed framework for DIM evaluation includes four steps: Firstly, square patches are detected from the ALS data and validated (Section 2.1); Secondly, corresponding DIM points are searched for each patch and the patches are further screened based on patch-based attributes (Section 2.2); Thirdly, quality measures are computed (Section 2.3); Finally, statistical analyses are performed on the valid patches.

2.1. Patch detection

The goal of patch detection is to localize candidate planar patches on the ALS point cloud. The patches taken as samples should be selected from the ALS data. The selection of patches should further avoid data gaps and breaklines. Planar patches of uniform size with acceptable noise level are considered valid and thus used for evaluation purpose. The examples used in this paper all make use of patches on the ground. The framework could, however, equally well be applied to planar non-ground patches.

In Fig. 1, a workflow is depicted for detecting ground patches. Firstly, ground points are identified from the ALS data using the method of (Axelsson, 2000). Then planar segments are extracted from ground points using a surface-based growing method (Vosselman, 2013). This approach employs the 3D Hough transform to detect seed surfaces. Then the nearby points are added to the surface if the distance from a certain point to the fitted plane is below a certain threshold. After new points are added to the segment, the plane parameters are recalculated before testing the next point. Slight over-segmentation is preferred over under-segmentation: over-segmentation can ensure better planarity and help avoiding breaklines in the segments.

After segmentation, the laser points with segment labels should be screened to discard small clusters or noisy segments. Features listed in

Table 1
Segment-based features for extracting smooth segments.

Feature	Description
Segment size	Number of points in the segment
linearity of segment	$(\lambda_1 - \lambda_2)/\lambda_1$, λ_1 is the maximum eigenvalue of the covariance matrix (Weinmann et al., 2015)
Plane slope	Normal direction of the fitted plane
Average angle	Mean of the angles between local point normals and the fitted plane normal
Residual of plane fitting (RPF)	Standard deviation of the distances between points and the plane fitted to the segment

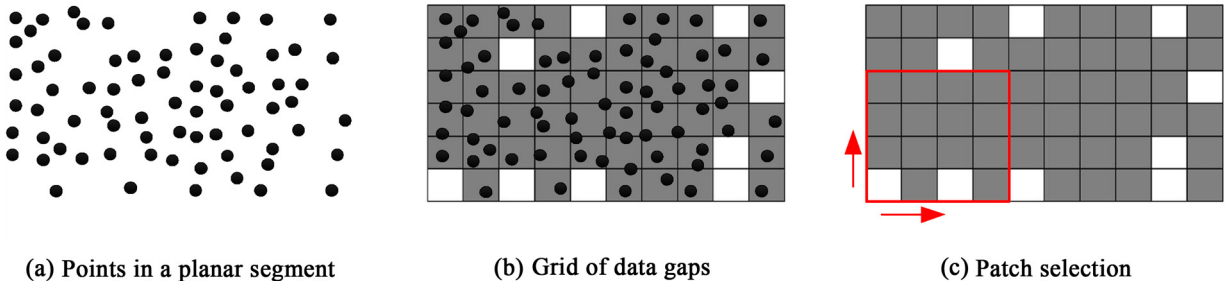


Fig. 2. Patch selection from the data gap grid in the horizontal space. In (b) and (c), the white cells indicate data gaps or empty cells, the grey cells indicate cells with points. (c) shows that the patch size is 4 × 4 cells represented by the red frame: the patch is valid only if there is no data gap in the 16 cells. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1 are used to remove these small or noisy segments. *Segment size* is used to eliminate small segments; *linearity of segment* is used to eliminate narrow segments; *Plane slope* is used to exclude segments on steep slopes; *average angle* and *residual of plane fitting* (*RPF*) are used to eliminate noisy clusters. A segment is kept only if it passes the check based on the five feature.

After smooth segments are obtained, patch selection is implemented in the bounding box of the segments. Fig. 2 shows that the bounding box is calculated around all the points in the segment. A raster grid is built within the bounding box in the horizontal space. If there is no point within a certain grid cell, the grid cell is set to *empty*, i.e. white cells in Fig. 2(b) and (c).

A patch is compiled out of several initial grid cells and is within the bounding box of the segment. The patch size should be scaled with the point density. It should be large enough to contain sufficient points but small enough to guarantee a large number of samples. In this paper, the cell size is set to 0.5 × 0.5 m and one patch contains 4 × 4 cells (see Fig. 2(c)). Hence, the patch size is 2 m × 2 m. If no data gap is detected in any cell within this patch, this patch is valid. In this way, the patch selection method can automatically avoid the locations of data gaps in the segments. The cell size is determined according to the laser point density. It should be large enough to guarantee at least one point in each cell in areas without data gaps. Additionally, in order to speed up the iteration, the stride can be two or more grid cells each time. Due to the "brute-force" search over dense grid cells, many selected patched are overlapping. The densely overlapping patches are screened automatically based on the spatial relationship to make sure that a certain location in the study area is used only once.

2.2. Patch selection

After patch detection, the candidate patch locations were obtained. The DIM points of a certain patch are selected according to the overlapping ALS patch. That is, the selection of DIM points adopts the same bounds as the ALS patch. Rule-based screening is implemented again at the patch level as previously implemented on the ALS segments in Section 2.1. Two types of rules are employed to select the patches. The first type of rules are to ensure a reliable quality analysis and check two patch properties:

(1) *Number of points* in the DIM patch: The DIM patches with data gaps are eliminated.
(2) *Mean deviation between ALS patch and DIM patch*: ALS data and aerial imagery could be captured at different times. This rule is to ensure that the *mean deviation* is caused by dense matching error but not by natural or man-made changes in between the ALS data and DIM data. The threshold for *mean deviations* should be for example at 0.99 quantile of the mean deviations.
We also calculate patch properties that allow us to distinguish between the quality of DIM points in various types of terrain:
(3) *Shading attribute*: The dense matching points in shadow often contain blunders and artefacts. For example, the dense matching errors along narrow alleys (often in shadow) are supposed to be much larger than the errors in the open area. The shadow mask is calculated from an orthoimage based on a grayscale histogram (Sirmacek and Unsalan, 2009). Only if all the four corners and the center location of a certain patch lie in the non-shaded area, the patch is accepted as non-shaded patch. Only if all the five locations are located in shadow, this patch is shaded patch.
(4) (4) *Green index*: DIM points on grassland will also have a different quality from DIM points on bare earth. The reason is that dense matching usually delivers points on top surface of grass, while laser scanning can penetrate the grass and represent the soil surface. In this case, the computed mean deviations will contain not only dense matching errors, but also the grass height. The Normalized Excessive Green Index (nEGI) in Eq. (1) is used to determine the vegetation patches on the orthoimages (Qin, 2014).

$$nEGI = (2G - R - B)/(2G + R + B) \quad (1)$$

Similar to the shading attribute, only if all the four corners and the center of a patch are labeled as vegetation, this patch will be considered as vegetation patch. After patch selection, DIM evaluation and statistical analysis can be performed based on the valid patches for different types of terrain.

2.3. Patch-based quality measures

The quality measures are calculated at each patch. This paper evaluates two factors related to the data quality:

– Accuracy: the deviation between the compared data and the ground truth (or reference data).
– Precision: the relative closeness of many measurements (in our case, dense matching points) to each other, i.e. the level of random noise.

Accuracy and precision are independent of each other. This paper only focuses on the vertical component of the point clouds and DSMs. Assuming that the 3D data show a normal distribution and contain no blunders, Table 2 shows the quality measures calculated at the patch level and the photogrammetric block level to represent the data accuracy and precision. The patch-based measures are aggregated into the block-level quality measures. In Table 2, i denotes the index of a patch in the whole block; j denotes the index of a specific DIM point in a certain patch. Δh_{ij} denotes the deviation from the jth DIM point to the plane which is fitted to all the ALS points within the ith patch. n_i denotes the number of DIM points in the ith patch. m denotes the number of patches in the whole block which is also the sample size for statistical analysis.

In Table 2, *Mean deviation* and *Standard deviation* are to indicate accuracy and precision at single patch level (Index i indicates a patch). Both $\overline{\mu}$ and σ_μ are measures indicating the accuracy at the block level. Specifically, a larger σ_μ indicates more dispersed patch-based errors in the block. In addition, μ_σ is to indicate the level of precision in the whole block.

Table 2
Quantitative quality measures for dense matching evaluation.

Level	Quality Measure	Definition	Meaning
Patch-level	*Mean deviation*	$\mu_i = \frac{1}{n_i}\sum_{n_i}^{j=1} \Delta h_{ij}$	Accuracy of DIM points in a certain patch w.r.t. the reference ALS plane
	Standard deviation	$\sigma_i = \sqrt{\frac{1}{n_i - 1}\sum_{n}^{j=1} (\Delta h_{ij} - \mu_i)^2}$	Precision of DIM data in a certain patch
Block-level	*Mean of mean deviations*	$\overline{\mu} = \frac{1}{m}\sum_{m}^{i=1} \mu_i$	Overall accuracy of DIM data in the whole block
	Standard deviation of mean deviations	$\sigma_\mu = \sqrt{\frac{1}{m-1}\sum_{m}^{i=1} (\mu_i - \overline{\mu})^2}$	Variation of accuracy measures in the block
	Average standard deviation	$\mu_\sigma = \sqrt{\frac{1}{m}\sum_{m}^{i=1} \sigma_i^2}$	Overall precision of DIM data in the block

In addition to the point cloud from dense matching, a DSM is also obtained from a standard photogrammetric workflow. The DSMs from a photogrammetric workflow can be in grid data structure but saved in point cloud format. Same with cropping point cloud patches, the DSM patches are generated by cutting out the corresponding patch area from the raster DSM.

3. Study area and experimental setup

3.1. Study area

The study area is located in Enschede, The Netherlands. Fig. 3 shows the dense matching block and the area for quality evaluation (1.6 km^2). This area is a densely-built urban area mainly covered by buildings, roads, squares, railways and vegetation. 510 aerial images including 102 nadir images and 408 oblique images were obtained by *Slagboom en Peeters* in 2011 together with exterior orientations. The tilt angle of oblique view is approximately 45°. The image size is 5616 × 3744 pixels. The GSD of nadir images equals 0.1 m. The overlap of nadir images is approximately 75% both along track and across track. The ALS data were acquired in 2007. The standard deviation of height differences between overlapping strips was around 2 cm (unpublished side result of the analyses in (Vosselman, 2008)). The absolute height accuracy has not been analyzed before. In the block, 105 ground reference targets (RTs) were measured with a Leica CS15 receiver using real time kinematic GPS. When collecting RTs, the accuracy of almost all the 105 RTs was better than 0.02 m in X, Y and Z directions, respectively; For several RTs, however, the accuracy in one or two directions were in between 0.02 m and 0.03 m. All of the RTs were the corners of zebra crossings, centers of manholes or other distinctive corners in the urban scene.

The RTs are used to evaluate the ALS quality. Since all the RTs are located in the open area, planes are fitted to the neighboring ALS points. The vertical residual from a RT to the fitted ALS plane is calculated as the indicator for the ALS accuracy. Results show that the mean deviation (μ) and standard deviation (σ) between the RTs and the fitted ALS plane are 0.013 m and 0.031 m. Furthermore, if the residual from RT to the ALS fitted plane is larger than three times of the standard deviations (σ), this RT will be discarded. This cross-verification ensures that both the ALS data and RTs used in the BBA, dense matching and DIM evaluation are reliable. Finally, 99 RTs passing this cross-verification are used as GCPs or check points in the BBA.

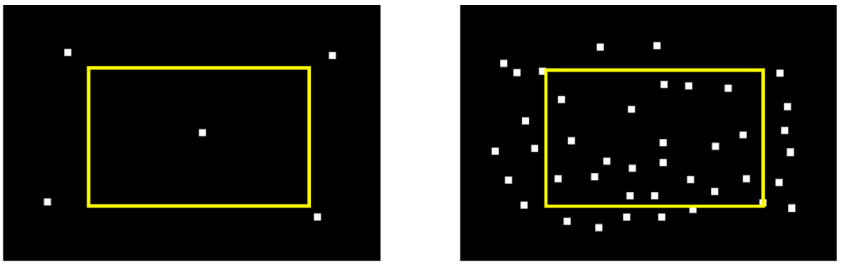
3.2. Bundle adjustment (BBA)

In the step of BBA, two configurations with 5 and 44 GCPs are set up for comparative study. The motivation to use and evaluate 2 different GCP-scenarios is to check whether block deformation, possibly caused by an insufficient GCP distribution, or by overfitting effects, will be observed by our evaluation method. The GCPs are evenly distributed in the block in both scenarios (Fig. 4). When 5 or 44 RTs are used as GCPs, the remaining 94 and 55 RTs are taken as check points, respectively. Note that direct sensor orientation elements available in this dataset are considered unreliable, therefore an indirect senor orientation approach is implemented. The results of the two configurations with 5 GCPs and 44 GCPs are presented in Sections 4.1 and 4.2, respectively.

The BBA was run in Pix4Dmapper Pro (version 3.2) on the original full resolution images. The standard deviation of the 3D GCPs was set to 0.02 m (default value in Pix4D) which controlled the GCPs weights in BBA. Table 3 shows the vertical RMSEs at GCPs and check points (CPs). When the number of GCPs increases from 5 to 44, the BBA network becomes more difficult to fit. Hence, the RMSE at GCPs increases. Meanwhile, the overall BBA accuracy is improved which is supported by the lower RMSE at check points.

(a) Dense matching points of the whole block (b) Orthoimage of the area for quality evaluation

Fig. 3. Dense matching block and orthoimage for the evaluation area. The area in the yellow frame in (a) corresponds to the area shown in (b). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a)Configuration wih 5 GCPs (b)Configuration with 44 GCPs

Fig. 4. Two configurations with different numbers of GCPs used in bundle adjustment. The white dots show the GCP distributions in the block. The yellow rectangle indicates the area for DIM evaluation (1.6 km^2). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Vertical RMSEs at GCPs and CPs when horizontal accuracy of GCPs is set to 0.02 m.

Number of GCPs	Number of CPs	RMSE at GCPs (m)	RMSE at CPs (m)
5	94	0.007	0.060
44	55	0.018	0.031

3.3. Dense image matching

For the execution of dense image matching, we select the state-of-the-art software SURE (Surface Reconstruction, version 2.1.0.33) from nFrames. A few researches have reported its performance in data accuracy (Haala and Rothermel, 2012; Rothermel et al., 2012; Ressl et al., 2016). The dense matching algorithm in SURE is a tube-shaped SGM (t-SGM). The SGM method in (Hirschmüller, 2008) is improved by restricting the disparity searching space which leads to a higher efficiency. Furthermore, the redundant disparity information is exploited to eliminate blunders and increase the accuracy of depth.

The interior orientation (IOs) and exterior orientation (EOs) elements are imported from Pix4D. Several parameters are supposed to control the dense matching quality. *Minimum Model Count* (*MMC*) represents the minimum number of models for a 3D point to be considered valid during triangulation. A larger *MMC* increases the reliability of generated points but also leads to a lower number of accepted matched points. When *MMC* is set large (e.g. ≥ 3), we find that many data gaps appear in narrow alleys. Hence, *MMC* is fixed to 2 in all our experiments. The image scale for dense matching is fixed to 1/2 so that the dense matching pipeline can be much faster compared to running at full scale. Note that different image scales used in dense matching will also affect the dense matching quality, but the impact of image scale is not the focus of our paper. The interpolation method for DSM generation is set to *Inverse Distance Weighting* (IDW). The resolution of the DSM grid is 0.1 m, i.e. equal to the size of nadir GSD.

The DIM data quality based on the configuration with 5 GCPs is evaluated to study the two issues: (1) The impact of the additional use of oblique images on the dense matching accuracy and precision; (2) Whether the accuracy of point cloud and DSM from a photogrammetric pipeline are the same. In summary, four data sets are obtained:

(1) GCP05_N + O_PC; (2) GCP05_N + O_DSM;
(3) GCP05_N_PC; (4) GCP05_N_DSM.

The naming scheme of the four data sets above shows different parameters of the data. GCP05 means 5 GCPs are used in BBA; N + O indicates that both nadir (N) and oblique (O) images are used in dense matching; "N" indicates that only nadir images are used in dense matching; PC or DSM refers to point cloud or DSM, respectively.

3.4. Parameter settings for DIM evaluation

In the segmentation step during patch detection, a surface growing radius of 1.0 m and maximum distance between point and fitted plane of 0.2 m are employed according to the point cloud density and noise level (Vosselman, 2013). The thresholds for the rules in Table 1 are set based on 200 valid segments and 200 invalid segments: *segment size* is 100, *linearity of segment* is 0.99, *plane slope* is 45°, *RPF* is 0.1 m, and *average angle* is 5°. The histograms of each feature for valid and invalid segments are depicted, respectively. Then the value that can best separate the two groups of segments is manually taken as the threshold. For example, the *segment size* is set according to the histogram of point amounts in these 200 valid segments. The smallest segment size is 100. Also segments with less than 100 points are very likely to be small noisy clusters. The segments with *linearity of segment* value larger than 0.99 are likely to be poles or other linear structures according to the histograms of invalid segments.

In patch screening, the threshold for *number of points* is determined from the histograms of the number of points in the valid and invalid DIM patches. The *mean deviation* threshold is set by adding the 0.99 quantile of the mean deviations with some small tolerance value (e.g. 0.02 m). The threshold for nEGI in Eq. (1) is set to 0.1 to recognize vegetation (Qin, 2014).

4. Experimental results

4.1. Results of the configuration with 5 GCPs

After patch selection, 7391 patches on the grassland, 2111 patches in the shadow and 24,634 non-shaded patches without vegetation were extracted. In the first analyses, only non-shaded patches without vegetation are evaluated. Fig. 5 shows some examples of these patches marked in blue, which are further used for DIM evaluation. Specifically, the left figure shows the selected patches on the central bus station of Enschede. The white stripes are actually platforms higher than the grey ground by around 0.2 m. The proposed algorithm performs well in extracting planar patches away from breaklines.

The patches of 2 m × 2 m are selected in the whole block, i.e. 0.1 km^2 in total. In order to make the block-level quality measures comparable, the same patch samples are used to evaluate the four data sets. Table 4 shows the quality measures at the block level calculated for the four data sets.

4.1.1. Evaluation of the impact of oblique images

The first row of Table 4 shows the comparison between GCP05_N + O_PC and GCP05_N_PC. A general finding is that when both nadir and oblique images are used in dense matching, all the three quality measures are better than the measures of configuration with only nadir images. The $\overline{\mu}$ improves remarkably by 0.014 m from 0.016 m to 0.002 m when oblique images are used; The σ_μ improves very slightly by 0.005 m; The μ_σ also improves by 0.012 m.

The distribution of mean deviations in the whole block for the two configurations are shown in Fig. 6. A normal distribution is estimated using the mean and standard deviation calculated from the same data set. The normal distribution is scaled and then superimposed on the histogram to visualize the deviation between the real measurements and a normal distribution (Höhle and Höhle, 2009). In each histogram, the horizontal axis indicates the patch-based mean deviation, the

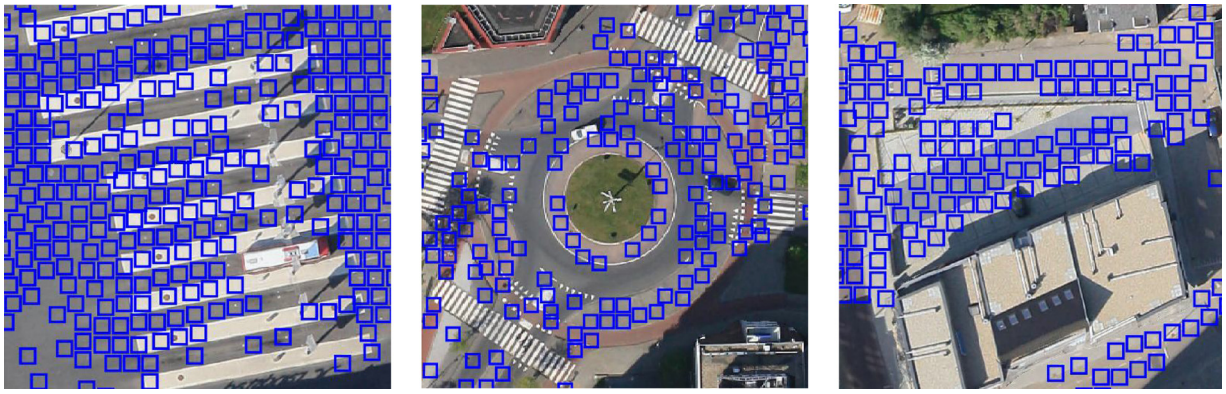


Fig. 5. Examples for extracted patches marked by blue squares. Patch size is 2 m × 2 m. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Quality measures for point clouds and DSMs in configurations with 5 GCPs (Unit: m).

Data sets	N + O			N		
	$\overline{\mu}$	σ_μ	μ_σ	$\overline{\mu}$	σ_μ	μ_σ
Point cloud (PC)	0.002	0.040	0.094	0.016	0.045	0.106
DSM	0.034	0.060	0.048	0.024	0.066	0.083

vertical axis indicates the frequency of patches in the whole block. The scale and interval of the axes for the two histograms are all the same.

The peak of Fig. 6(a) is located at approximately 0 which corresponds with $\overline{\mu} = 0.002$ m in Table 4. The histogram is centralized and "thin" in shape which corresponds with $\sigma_\mu = 0.040$ m. The mean deviations range from −0.060 m to 0.070 m which means that in most patches, the vertical error of dense matching is better than 1 GSD. Fig. 6(b) shows a relatively dispersed histogram compared to Fig. 6(a). In Fig. 6(b), the peak is located at 0.016 m at the horizontal axis. The range of mean deviations from −0.060 m to 0.090 m is slightly wider than in Fig. 6(a).

Fig. 7 shows the patch-based mean deviations in the block for the data GCP05_N + O_PC colored according to the absolute mean deviation values. That is, each pixel indicates a patch location. The patch samples are densely distributed in the whole block, mainly on roads, squares and parking lots. According to the color bar, the absolute mean deviations range from 0 to 0.12 m. Generally, the dense matching errors are homogenous in the whole block. However, in some locations, especially along narrow alleys, the mean deviations may get worse. The

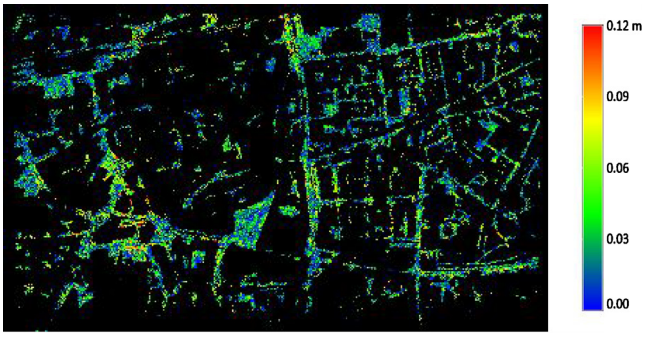


Fig. 7. Patch-based mean deviations in the whole block colored by the absolute values for the data GCP05_N + O_PC. Color coding from blue to red indicates that the mean deviation increases. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

point clouds in those regions get less accurate for two reasons: First, there are usually less visible image rays on the ground; Second, the image contrast is poor so dense matching will be problematic when finding correspondences among images.

4.1.2. Visualization of patch-based mean deviations

The patch-based mean deviations are visualized in Fig. 8. This square paved by concrete in our study area is relatively smooth. The square patches are colored based on positive or negative values. In Fig. 8(b) and (d), the mean deviation values are filled in the squares to visualize the dense matching errors in each patch.

Fig. 8(a) and (c) shows that the patch-based mean deviations vary

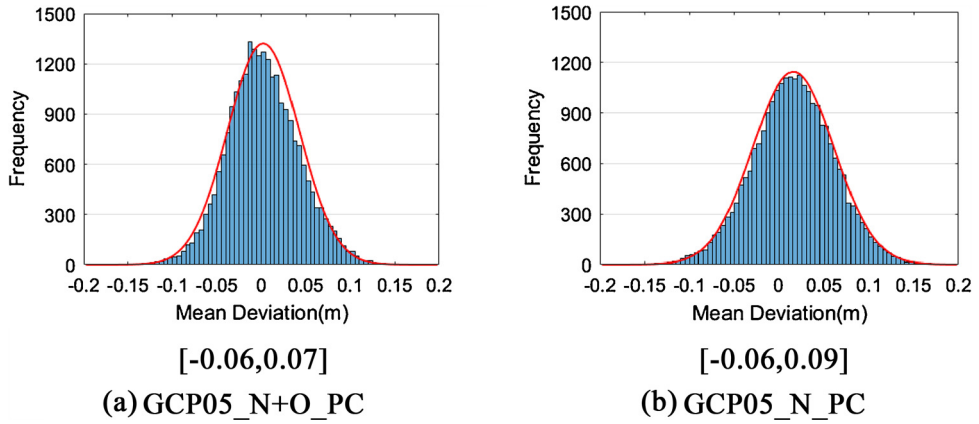


Fig. 6. Distribution of mean deviations for 24,634 non-shaded ground patches (Unit: m). All the histograms are overlaid with an estimated normal distribution. The interval below each histogram refers to the 0.05 and 0.95 quantile, respectively.

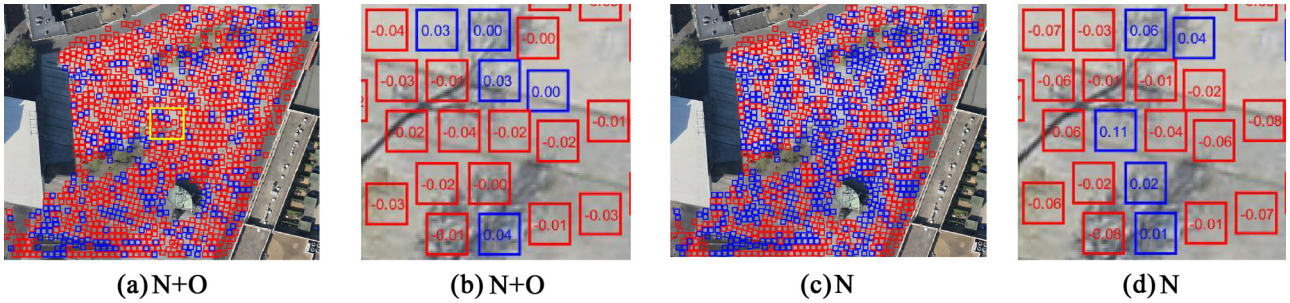


Fig. 8. Visualization of patch-based mean deviations for the data sets GCP05_PC. The blue patches indicate positive values while the red indicates negative values, (a) and (b) show the mean deviations for the data GCP05_N + O_PC, (c) and (d) show the quality measures for GCP05_N_PC. The yellow rectangle on (a) is zoomed in and filled in the mean deviation values as shown in (b) and (d). (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.).

between positive and negative on the square. The positive value indicates that the point cloud surface from dense matching is higher than the point cloud surface from laser scanning, and vice versa. We can infer that the point cloud surface from dense matching is fluctuating around the referred ALS surface. In addition, comparing Fig. 8(a) and (c) in the red and blue patterns, or (b) and (d) in the values shows that whether or not oblique images are used in dense matching makes a large difference on the local accuracy.

4.1.3. Dense matching quality on different terrain types

The profiles of ALS points and DIM points are shown in Fig. 9. All the DIM data are generated from the configurations of 5 GCPs. The profile interval in the horizontal space is 0.25 m. Fig. 9(a) shows the profiles along a smooth downtown square. Both profile N + O and profile N are fluctuating around the ALS profile. As a comparison, Fig. 9(b) and (c) shows the profiles across a narrow shaded alley and short grass. The deviation between ALS profile and DIM profile in Fig. 9(a) and (b) is caused by the dense matching errors on the smooth surface with poor texture while the deviation in Fig. 9(c) is mainly caused by the rugged grassland surface itself.

As quantitative analysis, the quality measures for three different land cover types: open ground without vegetation, grassland and shadow are shown in Fig. 10. Statistical measures reveal no significant difference between open sealed terrain and grassland. This is likely caused by the very low height of the vegetation. The accuracy in shaded area is lower than the accuracy on sealed ground and grassland.

Concerning the μ_σ (see the green broken-line in Fig. 10), the noise level of DIM data in shaded areas is higher than the noise level in open sealed terrain and grassland. Interestingly, the μ_σ value on grassland (0.085 m) is lower than open sealed terrain (0.094 m). Two factors are influencing the noise level: the contrast of the object surface and the roughness of the object surface. The concrete ground in our study area is mainly characterized by smooth and textureless surface. Dense matching between two images with little texture will be ambiguous and some repetitive patterns will appear in the point cloud surface (see Fig. 9(a)). In contrast, the grassland is covered with low grass with rich texture. Therefore, the noise level on open sealed ground caused by dense matching seems to exceed the noise level on the low grassland.

4.1.4. Comparison between point cloud and DSM

As expected, the μ_σ of DSMs in Table 4 indicates that the noise level is much lower than within the point clouds since interpolation is

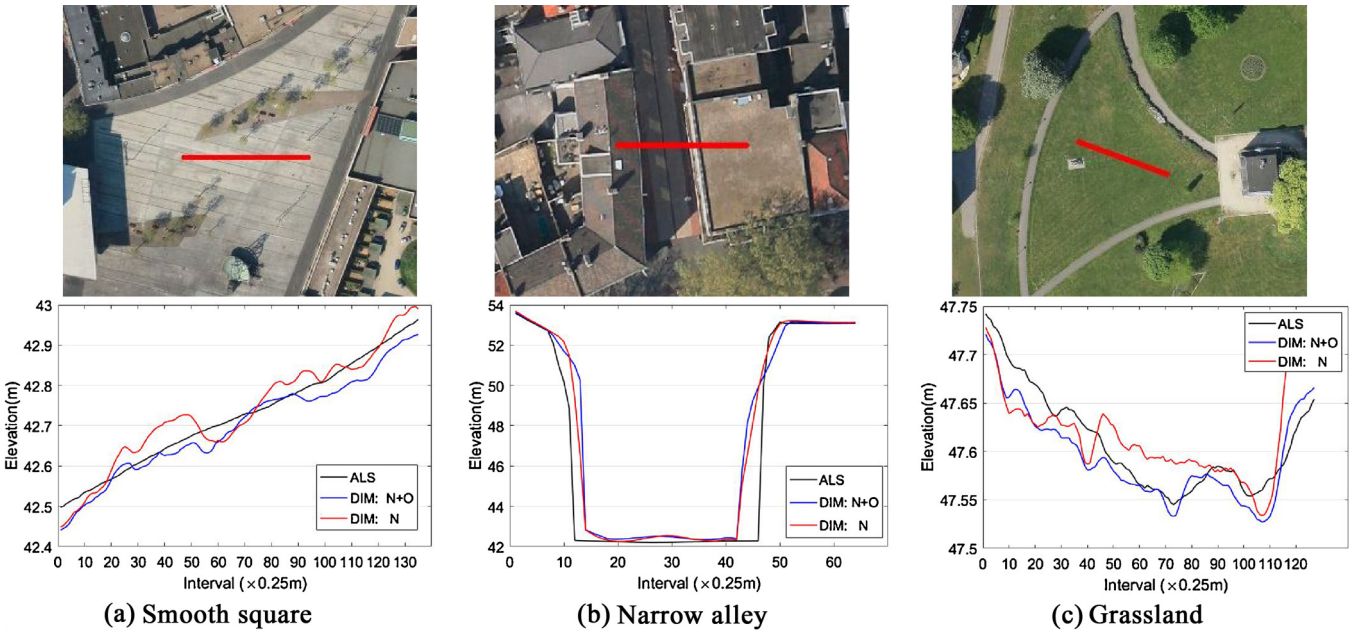


Fig. 9. 3D data profiles for three area: (a) smooth concrete square, (b) narrow alley and (c) grassland. The top row shows the orthoimages with profiles marked in red. The bottom row depicts the relevant profiles for ALS point cloud (black), DIM_N + O (blue) and DIM_N (red). The horizontal axis indicates the interval along the profile, the vertical axis indicates the elevation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

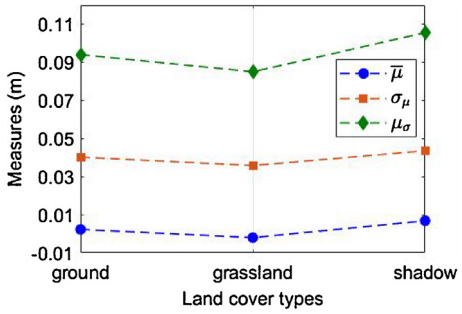


Fig. 10. DIM quality on three terrain types for the data GCP05_N + O_PC. The horizontal axis indicates three terrain types: open ground without vegetation (label "ground"), grassland, shadow. The vertical axis indicates quality measures. The three broken-lines indicate $\bar{\mu}$, σ_μ and μ_σ, respectively.

employed. Based on our evaluation framework, we observe a bias between point clouds and DSMs from the software pipeline. The first column in Table 4 shows that the difference of $\bar{\mu}$ between point cloud and DSM for N + O is 0.032 m. The second column in Table 4 shows that the DSM surface is higher than the point cloud surface by 0.008 m when only nadir images are used in dense matching. Therefore, we conclude that the interpolation process changes the data accuracy; and the magnitude seems to depend on the point cloud density. Mandlburger et al. (2017) also reported the same deviation between point cloud surface and DSM surface so this deviation might be caused by the interpolation process in the SURE software. In order to visually analyse the deviation between point clouds and DSMs, refer to Fig. 11.

Fig. 11(a) shows a clear deviation between the peaks of the two histograms while in Fig. 11(b) the deviation is relatively smaller (0.032 m vs. 0.008 m) when oblique images are not used. Another finding is that the distribution of mean deviations for DSMs is more dispersed than for the point cloud, which corresponds with Table 4 that the σ_μ of DSMs is larger than point clouds. In summary, although the noise level is reduced from point clouds to DSMs, the absolute accuracy is changed during interpolation and the error distribution in DSMs is more dispersed than in point clouds.

4.2. Impact of number of GCPs and weights

We evaluate the point cloud *GCP44_N + O_PC* using the same 24,634 patches as we did for the configuration with 5 GCPs. The calculated quality measures are listed in the last row of Table 5. It is obvious that the dense matching accuracy indicated by $\bar{\mu}$ and σ_μ gets worse when more GCPs are used.

However, Table 3 in Section 3.2 shows that when the number of GCPs increases from 5 to 44, the RMSE at check points in BBA reported by Pix4D decreases from 0.060 m to 0.031 m. That is, the check points indicate that the accuracy of BBA is getting better when more GCPs are used. Empirical knowledge from previous studies (e.g. Gerke et al., 2016) also indicates that the more GCPs, the better the BBA accuracy will be.

Therefore, when the number of GCPs increases from 5 to 44, the accuracy of BBA gets improved but the accuracy of the DIM point cloud deteriorates. This contradictory finding is further evaluated by visualizing the patch-based mean deviations in Fig. 12.

Fig. 12 shows inhomogeneous error distribution in the block. Even though the absolute mean deviations still range from 0 to 0.12 m, the absolute mean deviations in the southwest of the block is generally larger than the other parts. The block is thus verified to be overfitted. The image block with large forward and sideward overlaps results in a very strong network of bundles. When many GCPs are used with relatively high weights compared to the tie points, the noise in the GCPs leads to a deformation of the network of bundles. The resulting errors in the exterior orientations propagates to locally systematic errors in the dense matching point cloud.

In order to check whether the BBA network is overfitted, the a priori standard deviation of the GCPs is set to 0.05 m in the BBA in Pix4D.[1] In this case, the bundle adjustment network controlled by the GCPs gets "loose". The BBA result is that the RMSE at GCPs is 0.019 m and the RMSE at check points is 0.031 m. Compared with Table 3, the RMSEs at GCPs and check points change very slightly. Then we evaluate the new point cloud generated by SURE with new orientations. We observe a large improvement in the point cloud quality. The $\bar{\mu}$, σ_μ and μ_σ for the new point cloud (from GCP standard deviations of 0.05 m) is 0.011 m, 0.044 m and 0.094 m, respectively. Comparing these results to those obtained with the higher GCP weights (Table 5), in particular the $\bar{\mu}$ value of 0.011 m indicates that the systematic error is strongly reduced.

The overfitting effect is alleviated as shown in Fig. 13. The homogeneity level of mean deviations in the block is much better than Fig. 12 and no remarkable systematic deviations appear.

5. Discussion

In our evaluation framework, both the $\bar{\mu}$ and σ_μ are used to represent the dense matching accuracy in the block. The $\bar{\mu}$ indicates the general bias of the DIM points from the reference while the σ_μ indicates the dispersion level of the dense matching errors. In Table 4, when oblique images and nadir images are both used in dense matching, the $\bar{\mu}$ gets improved, but the σ_μ keeps relatively stable.

Even when the ALS accuracy is verified in Section 3.1, the noise in the ALS data still has a small impact on the computed quality measures. These should be taken into account when assessing whether the quality of a DIM point cloud meets the requirements of a project.

A good point cloud should not only be accurate but also represent the object details with little random noise. When tuning the parameters in dense matching, the key is to balance between data gap level and noise level. Dense matching quality depends largely on the image contrast and texture. In order to obtain less noisy points from the SURE software on the problematic locations (e.g. narrow streets or in shadow), the parameter *MMC* should be set as large as possible as long as the data gap level is still acceptable. The dense matching can be more challenging in densely-built urban areas. The dense matching quality on open smooth ground with better texture is usually more reliable than locations with poor texture. The DIM data profiles in Fig. 9 show that dense matching will be problematic in representing the ground details along narrow alleys.

Concerning the overfitting in the BBA, this effect cannot be detected by evaluating the BBA accuracy based on the RMSEs determined with a few check points as common in many previous studies, but can be detected in our evaluation framework. Our finding shows that the RMSEs of check points in the BBA are not equivalent to the point cloud accuracy from dense matching. In our two comparative experiments, the BBA network becomes overfitted when 44 GCPs with high weights are employed in BBA; In contrast, the point cloud GCP05_N + O_PC with only 5 GCPs has already achieved the accuracy better than 1 GSD. It should, however, be noted that when only few GCPs are used, the BBA may become more sensitive to the selection of GCPs. That is, the BBA network is easier to become biased due to one or two inaccurate GCPs.

6. Conclusion

In order to check the potential of using point clouds derived by airborne dense matching as effective alternatives to airborne laser scanning data, we have presented a framework for evaluating the quality of 3D point clouds and DSMs generated by dense image matching in urban area. Square patches of uniform size are extracted

[1] It would be more appropriate to lower the a priori standard deviation of the tie points, but Pix4D does not allow to set the tie point standard deviation.

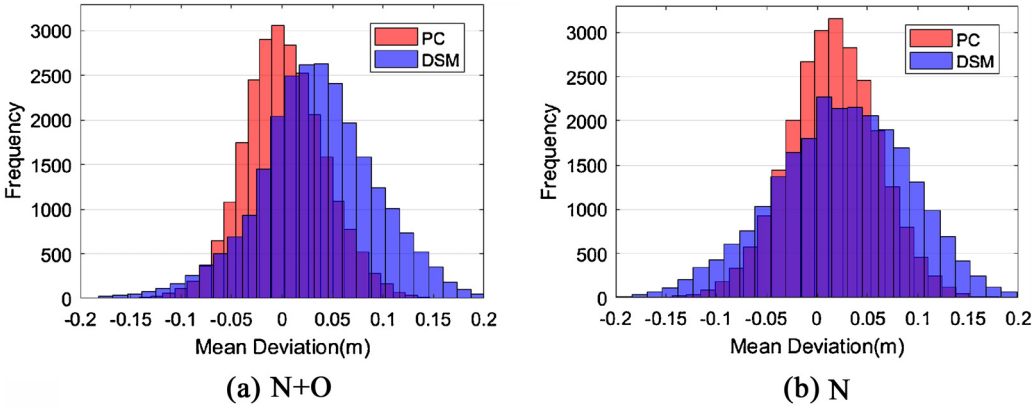


Fig. 11. Overlaid histograms of mean deviations for point cloud and DSM. Red histograms indicate the distributions of point clouds; blue histograms indicate the distributions of DSMs. "PC" in the legend indicates point cloud. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.).

Table 5
Quality measures for point clouds when the GCP weights in BBA is set to 0.02 m (Unit: m).

Configuration	$\bar{\mu}$	σ_μ	μ_σ
GCP05_N + O_PC	0.002	0.040	0.094
GCP44_N + O_PC	−0.026	0.049	0.098

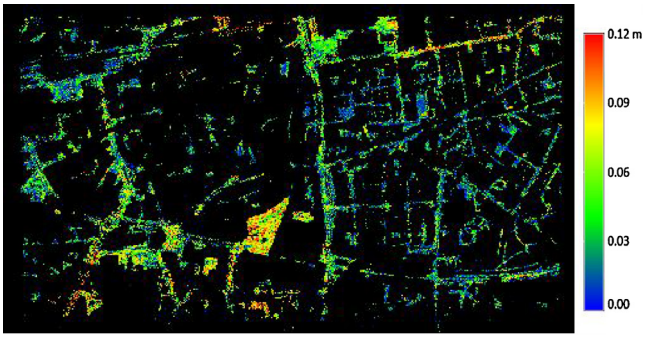


Fig. 12. Distribution of mean deviations (absolute values) for GCP44_N + O_PC when GCP weight for BBA is set to 0.02 m.

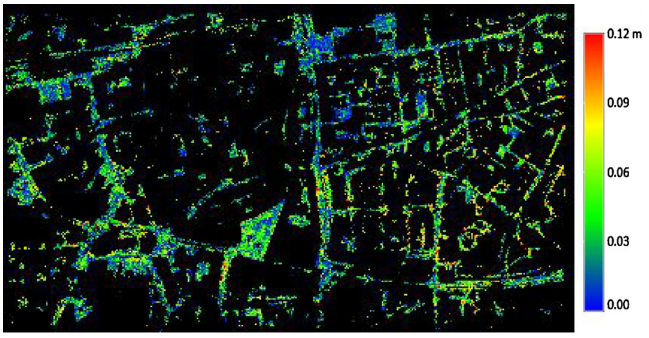


Fig. 13. Distribution of mean deviations (absolute values) for GCP44_N + O_PC when GCP weight for BBA is set to 0.05 m.

from planar terrain with the guidance of ALS data. The previous evaluation work based on check points simply reveals the BBA accuracy, which is not equivalent to the accuracy of photogrammetric point clouds. In contrast, our evaluation framework based on large sample size is able to reveal the distribution of dense matching errors in a whole photogrammetric block. This framework based on "plane-to-plane" distance is robust to possible blunders and artefacts in the DIM points. Robust quality measures are proposed to represent the dense matching accuracy and precision quantitatively. Experiments show that the optimal accuracy of DIM point cloud is as follows: the overall mean offset to the reference is 0.1 GSD; the maximum mean deviation reaches 1.0 GSD.

In order to further test the usability of the proposed framework, some factors that may affect the DIM quality are studied. Based on our evaluation framework, we find that when oblique images are used in dense matching together with nadir images, the accuracy of DIM point cloud improves and the noise level decreases on smooth ground areas. Concerning the DIM quality on different terrain types, the accuracy and precision on grassland is similar to those on open sealed ground in case of very low grass. Both the accuracy and precision level in shadow are worse than for the other two terrain types. The evaluation framework also reports a deviation between the point cloud and DSM generated by a single photogrammetric workflow. The deviation is less distinct when the point cloud density drops. When many GCPs with high weights are employed in BBA, the BBA network may become overfitted, which is reflected in the inhomogeneous distribution of the patch-based DIM errors. This problem cannot be detected by check points in BBA. While this paper evaluates the impact of oblique images and compares the point clouds and DSMs, future work can still study the impact of other factors (e.g. number of GCPs and their distribution, image scale, and *MMC*) on the DIM data quality.

References

Axelsson, P., 2000. DEM generation from laser scanner data using adaptive TIN models. Int. Arch. Photogram. Remote Sens. Spatial Inf. Sci. 33, 110–117.
Baltsavias, E.P., 1999. A comparison between photogrammetry and laser scanning. ISPRS J. Photogramm. Remote Sens. 54 (2), 83–94.
Cavegn, S., Haala, N., Nebiker, S., Rothermel, M., Tutzauer, P., 2014. Benchmarking high density image matching for oblique airborne imagery. Int. Arch. Photogram. Remote Sens. Spatial Inf. Sci. 40 (3), 45–52.
Dong, W., Lan, J., Liang, S., Yao, W., Zhan, Z., 2017. Selection of LiDAR geometric features with adaptive neighborhood size for urban land cover classification. Int. J. Appl. Earth Obs. Geoinf. 60, 99–110.
Gehrke, S., Morin, K., Downey, M., Boehrer, N., Fuchs, T., 2010. Semi-global matching: an alternative to LIDAR for DSM generation. Proceedings of the 2010 Canadian Geomatics Conf. and Symp. of Commission I 1–6.
Gerke, M., Nex, F., Remondino, F., Jacobsen, K., Kremer, J., Karel, W., Huf, H., Ostrowski, W., 2016. Orientation of oblique airborne image sets-experiences from the ISPRS/EUROSDR benchmark on multi-platform photogrammetry. Int. Arch. Photogram. Remote Sens. Spatial Inf. Sci. 41, 185–191.
Höhle, J., Höhle, M., 2009. Accuracy assessment of digital elevation models by means of robust statistical methods. ISPRS J. Photogramm. Remote Sens. 64 (4), 398–406.
Haala, N., Rothermel, M., 2012. Dense multi-stereo matching for high quality digital elevation models. Photogrammetrie-Fernerkundung-Geoinformation 4, 331–343.
Haala, N., Hastedt, H., Wolf, K., Ressl, C., Baltrusch, S., 2010. Digital photogrammetric camera evaluation–generation of digital elevation models. Photogrammetrie-Fernerkundung-Geoinformation 2, 99–115.
Hirschmüller, H., 2008. Stereo processing by semiglobal matching and mutual information. IEEE Trans. Pattern Anal. Mach. Intell. 30 (2), 328–341.
Hobi, M.L., Ginzler, C., 2012. Accuracy assessment of digital surface models based on WorldView-2 and ADS80 stereo remote sensing data. Sensors 12 (5), 6347–6368.
Jaud, M., Passot, S., Le Bivic, R., Delacourt, C., Grandjean, P., Le Dantec, N., 2016. Assessing the accuracy of high resolution digital surface models computed by PhotoScan® and MicMac® in sub-optimal survey conditions. Remote Sens. 8 (6), 465–482.
Kraus, K., Karel, W., Briese, C., Mandlburger, G., 2006. Local accuracy measures for

digital terrain models. Photogramm. Rec. 21 (116), 342–354.
Leberl, F., Irschara, A., Pock, T., Meixner, P., Gruber, M., Scholz, S., Wiechert, A., 2010. Point clouds: lidar versus 3D vision. Photogramm. Eng. Remote Sens. 76 (10), 1123–1134.
Maltezos, E., Kyrkou, A., Ioannidis, C., 2016. LiDAR vs dense image matching point clouds in complex urban scenes. Proc. SPIE 9688. Fourth Int. Conf. on Remote Sens. and Geoinform. of the Environ 9688, 1–10.
Mandlburger, G., Wenzel, K., Spitzer, A., Haala, N., Glira, P., Pfeifer, N., 2017. Improved topographic models via concurrent airborne lidar and dense image matching. ISPRS Ann. Photogram. Remote Sens. Spatial Inf. Sci. IV-2/W4 259–266.
Moussa, W., Wenzel, K., Rothermel, M., Abdel-Wahab, M., Fritsch, D., 2013. Complementing TLS point clouds by dense image matching. Int. J. Herit. Digit. Era 2 (3), 453–470.
Mura, M., McRoberts, R.E., Chirici, G., Marchetti, M., 2015. Estimating and mapping forest structural diversity using airborne laser scanning data. Remote Sens. Environ. 170, 133–142.
Nex, F., Gerke, M., Remondino, F., Przybilla, H.J., Bäumker, M., Zurhorst, A., 2015. ISPRS benchmark for multi-platform photogrammetry. ISPRS Ann. Photogram. Remote Sens. Spatial Inf. Sci. 2 (3), 135–142.
Nurminen, K., Karjalainen, M., Yu, X., Hyyppä, J., Honkavaara, E., 2013. Performance of dense digital surface models based on image matching in the estimation of plot-level forest variables. ISPRS J. Photogramm. Remote Sens. 83, 104–115.
Poon, J., Fraser, C.S., Chunsun, Z., Li, Z., Gruen, A., 2005. Quality assessment of digital surface models generated from IKONOS imagery. Photogramm. Rec. 20 (110), 162–171.
Qin, R., 2014. An object-based hierarchical method for change detection using unmanned aerial vehicle images. Remote Sens. 6 (9), 7911–7932.
Remondino, F., Spera, M.G., Nocerino, E., Menna, F., Nex, F., 2014. State of the art in high density image matching. Photogramm. Rec. 29 (146), 144–166.
Remondino, F., Nocerino, E., Toschi, I., Menna, F., 2017. A critical review of automated photogrammetric processing of large datasets. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 42, 591–599.
Ressl, C., Brockmann, H., Mandlburger, G., Pfeifer, N., 2016. Dense Image Matching vs. Airborne Laser Scanning–Comparison of two methods for deriving terrain models. PFG Photogrammetrie Fernerkundung Geoinformation 2, 57–73.
Rothermel, M., Haala, N., 2011. Potential of dense matching for the generation of high quality digital elevation models. ISPRS Hannover Workshop for High-Resolution Earth Imaging for Geospatial Information 331–343.
Rothermel, M., Wenzel, K., Fritsch, D., Haala, N., 2012. SURE: photogrammetric surface reconstruction from imagery. Proc. of LC3D Workshop 1–9.
Sirmacek, B., Unsalan, C., 2009. Damaged building detection in aerial images using shadow information. Recent Advances in Space Technologies 249–252.
Sofia, G., Bailly, J.S., Chehata, N., Tarolli, P., Levavasseur, F., 2016. Comparison of Pleiades and LiDAR digital elevation models for terraces detection in farmlands. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 9 (4), 1567–1576.
Tian, J., Schneider, T., Straub, C., Kugler, F., Reinartz, P., 2017. Exploring digital surface models from nine different sensors for forest monitoring and change detection. Remote Sens. 9 (3), 287–312.
Tomljenovic, I., Tiede, D., Blaschke, T., 2016. A building extraction approach for airborne laser scanner data utilizing the object based image analysis paradigm. Int. J. Appl. Earth Obs. Geoinf. 52, 137–148.
Toschi, I., Ramos, M.M., Nocerino, E., Menna, F., Remondino, F., Moe, K., Poli, D., Legat, K., Fassi, F., 2017. Oblique photogrammetry supporting 3d urban reconstruction of complex scenarios. Int. Arch. Photogram. Remote Sens. Spatial Inf. Sci. 42, 519–526.
Vosselman, G., 2008. Analysis of planimetric accuracy of airborne laser scanning surveys. ISPRS Int. Arch. Photogram. Remote Sens. Spat. Inf. Sci. 37 (3a), 99–104.
Vosselman, G., 2013. Point cloud segmentation for urban scene classification. ISPRS Int. Arch. Photogram. Remote Sens. Spat. Inf. Sci. 1, 257–262.
Weinmann, M., Jutzi, B., Hinz, S., Mallet, C., 2015. Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers. ISPRS J. Photogramm. Remote Sens. 105, 286–304.
Xiong, B., Jancosek, M., Elberink, S.O., Vosselman, G., 2015. Flexible building primitives for 3D building modeling. ISPRS J. Photogramm. Remote Sens. 101, 275–290.
Yang, B., Chen, C., 2015. Automatic registration of UAV-borne sequent images and LiDAR data. ISPRS J. Photogramm. Remote Sens. 101, 262–274.
Zhang, Z., Gerke, M., Peter, M., Yang, M.Y., Vosselman, G., 2017. Dense matching quality evaluation-an empirical study. IEEE Joint Urban Remote Sensing Event (JURSE) 1–4.