3D reconstruction from multi-view VHR-satellite images in MicMac

Ewelina Rupnik, Marc Pierrot-Deseilligny, Arthur Delorme

▶ **To cite this version:**

Ewelina Rupnik, Marc Pierrot-Deseilligny, Arthur Delorme. 3D reconstruction from multi-view VHR-satellite images in MicMac. ISPRS Journal of Photogrammetry and Remote Sensing, 2018, ⟨10.1016/j.isprsjprs.2018.03.016⟩. ⟨hal-02369304⟩

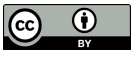
HAL Id: hal-02369304

https://hal.science/hal-02369304v1

Submitted on 18 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

ISPRS Journal of Photogrammetry and Remote Sensing 139 (2018) 201–211

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

3D reconstruction from multi-view VHR-satellite images in MicMac

Ewelina Rupnik [a,*], Marc Pierrot-Deseilligny [a], Arthur Delorme [b]

[a] LaSTIG, IGN, ENSG, Univ. Paris-Est F-94160, Saint-Mande, France
[b] Institut de Physique du Globe de Paris, Sorbonne Paris Cité, UMR 7154 CNRS, F-75005, Paris, France

ARTICLE INFO

Article history:
Received 19 September 2017
Received in revised form 29 January 2018
Accepted 13 March 2018
Available online 20 March 2018

Keywords:
VHR-satellite imagery
Multi-view
Bundle block adjustment
Dense image matching
Depth map fusion

ABSTRACT

This work addresses the generation of high quality digital surface models by fusing multiple depths maps calculated with the dense image matching method. The algorithm is adapted to very high resolution multi-view satellite images, and the main contributions of this work are in the multi-view fusion. The algorithm is insensitive to outliers, takes into account the matching quality indicators, handles non-correlated zones (e.g. occlusions), and is solved with a multi-directional dynamic programming approach. No geometric constraints (e.g. surface planarity) or auxiliary data in form of ground control points are required for its operation. Prior to the fusion procedures, the RPC geolocation parameters of all images are improved in a bundle block adjustment routine. The performance of the algorithm is evaluated on two VHR (Very High Resolution)-satellite image datasets (Pléiades, WorldView-3) revealing its good performance in reconstructing non-textured areas, repetitive patterns, and surface discontinuities.
© 2018 The Authors. Published by Elsevier B.V. on behalf of International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

The modern high resolution satellites are capable of frequent revisit times thanks to their pointing agility. Imagery collected only from nadir view is rich in details but has small ground footprints. The pointing agility overcomes this potential limitation by allowing the satellite to alter its look direction by a large angle and in a rapid manner. Other benefits are that, e.g., by adapting the imaging configuration the satellite can collect enhanced quality data by searching for cloudless conditions; or by providing rapid response in the event of natural or man-made disasters.

Examples of modern, well-established commercial satellites are the WorldView 1–4, Pléiades 1A/1B and SPOT 6/7 satellites. Their typical acquisition modes are (i) the consecutive imaging where the ground is captured continuously, or (ii) the target mode that samples the ground with non-contacting patches. Depending on the adopted acquisition mode, a satellite can view an area covering up to 1000×1000 km^2, all in a single pass.[1] This necessitates large viewing angles (up to $\approx 30^\circ$ in the standard mode, 45° in the extended mode but also 60° and more if desired) and manifests in both, less controlled base to height ratio (B/H), varying ground sampling distance (GSD), as well as possible multi-view acquisitions.

While the imaging configurations become more irregular, the state-of-the-art automated algorithms for image dense matching rely on a strong assumptions about visual resemblance of respective image patches Hirschmüller (2008); Pierrot-Deseilligny and Paparoditis (2006). The resemblance, however, is violated for wide-baseline, or diachronic acquisitions and the common dense matching similarity measures are not apt to effectively model such transformations. Paradoxically, from the standpoint of 3D reconstruction precision, the larger the B/H, the better the point intersection angles, hence the more precise coordinates' estimation.

The above, the increased availability, and the demand for versatile datasets motivates the research presented in this work. We lay down a 3D reconstruction pipeline focusing on a global 3D fusion algorithm, implemented in the free open-source software tool for photogrammetry – MicMac (Rupnik et al., 2017). The contributions are:

- unlike most current approaches applied to satellite datasets the approach is not limited to stereo-pairs processing (Reinartz et al., 2005; Karkee et al., 2008; Hirschmüller, 2008; Xu et al., 2010; Kuschk and dAngelo, 2013) but works with multi-view configurations of varying base lengths and resolutions;
- it is defined in 3D and handles occlusions (Section 3.2);
- it is insensitive to outliers and weighted by the image matching confidence indicator (Section 3.3.2);
- it is formulated as a discrete, semi-global optimization problem therefore circumvents the artefacts pertinent to local methods

* Corresponding author.
E-mail addresses: ewelina.rupnik@ign.fr (E. Rupnik), marc.pierrot-deseilligny@ensg.eu (M. Pierrot-Deseilligny), delorme@ipgp.fr (A. Delorme).

[1] www.intelligence-airbusds.com/www.satimagingcorp.com accessed 09/2017.

https://doi.org/10.1016/j.isprsjprs.2018.03.016
0924-2716/© 2018 The Authors. Published by Elsevier B.V. on behalf of International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS).
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

while being computationally and memory-use efficient (Section 3.3.3);
– no geometric constraints (e.g. surface planarity) are assumed and no auxiliary data (e.g. ground control points) are required.

The implemented methods are tested on numerous multi-view satellite image configurations using two datasets – a rural zone captured in a single epoch and an urban zone captured at various epochs across the year. All results are evaluated qualitatively and quantitatively by comparing them with ground truths of superior quality (only urban area).

The publication is organized as follows: Section 2 provides a concise literature review on state-of-the-art 3D fusion approaches; Section 3 addresses the reconstruction pipeline available in MicMac with the focus on the theoretical aspects of the employed fusion methodology; and finally, the experimental part is included in Section 4.

2. Related work

Research on depth map fusion (or integration) has a long history and varies across the applications and scales (i.e. terrestrial, aerial, satellite). Significant differences are in:

– local or global formulation of the fusion;
– description of the 3D point sample (e.g. including the spatial context, the scale attribute, a point's orientation, a quality indicator).

Local methods perform the fusion considering 3D sample points independently, or by taking into account a local context. As for the global methods, the fusion is defined over the entire scene as an energy functional. The energy typically contains the data term reflecting the confidence of a 3D point to belong to the surface, and the regularizing term which favours smooth surfaces. Inclusion of the *a priori* in form of a regularizing term handles the missing data and attenuates the noise.

Complex 3D objects. Local approaches. Early examples of surface fusion are local and based on filtering and averaging schemes. In Volumetric Range Image Processing (VRIP) by Curless and Levoy (1996) depths are accumulated in a voxel grid where each voxel is described by a weighted signed distance of that point to the nearest range surface along the line of sight of the sensor. The final surface is an isosurface, i.e. for each surface point its distance value is constant and equal to zero.

However, simple averaging is susceptible to outliers and smooths the surface's high frequency component. To overcome such artefacts Fuhrmann and Goesele (2011) use the notion of a point's scale and calculate a weighted average of the signed distance functions for points at similar resolutions. Kuhn et al. (2013) proposes an improved, hybrid version of both the VRIP and the work of Fuhrmann and Goesele (2011). The authors differentiate between changing 3D point's quality due to image acquisition geometry by adding a stochastic component to the distance function. Hence the fusion is rendered more rigorous. Despite being precise, all three approaches suffer from a large computational burden making them inappropriate for big datasets.

Complex 3D objects. Semi-global and global approaches. If no regularization is adopted such averaging techniques effect in noisy reconstructions or surface inconsistencies due to the mean distance field sign changes (Zach, 2008). Consequently, Zach et al. (2007) and Zach (2008) adopt a global variational approach. The data term is a signed distance function weighted with depth uncertainty and the regulariser is the first order total variation. As both methods work on a regular voxel grid kept in the memory, their scalability to large scenes is questionable (Ummenhofer and Brox, 2015). In contrast to Zach (2008), Pock et al. (2011) appoints the total generalized variational (TGV) model with a signed distance function describing the data and a second order smoothness term. It privileges piece-wise affine surface thus allows for reconstruction of slanted surfaces. Ummenhofer and Brox (2015) borrow the TGV formulation of the regularisation from Pock et al. (2011), and upgrade the distance function by the surface normal. The inclusion of the normal vector is said to better preserve surface discontinuities. A variational reconstruction and fusion was also studies by Vu et al. (2012). The pipeline consists of three main steps, i.e. restitution of the sparse point cloud cloud, initial mesh building while respecting the visibility constraint and the variational refinement. The variational energy is a function of a mesh-patch photo-consistency measure. The proposed sequence is suitable for large scene 3D reconstruction at reasonable processing times.

There exist alternative ways to describe 3D points by implicit functions. For instance, Papasaika et al. (2011) uses a sparse representation for plausible terrain shapes. Local digital elevation model (DEM) patches are represented as a weighted combination of basis terrain shapes defined over a local support, allowing to eliminate incoherent geometries from the fusion (Schindler et al., 2011). Kazhdan and Hoppe (2013) introduce the so-called indicator function which takes non-zero values only close to the surface. The algorithm accepts a set of oriented 3D samples as input, and seeks the indicator function that best approximates the points' orientations. Because it is formulated as a *Poisson* problem it provides a global solution while considering all the points at once. Small and big holes are filled, nonetheless, certain reconstruction artefacts are produced (e.g. merging of surface concave elements) as no acquisition modality, the point's scale and quality are considered. In Fuhrmann and Goesele (2011) the distance function is replaced with a sum of support basis functions. The 3D point's scale and orientations are the necessary input as the functions are derivatives of the Gaussian in the normal direction, parametrized by the samples' scale. Because it is solved locally it is has a good runtime performance, however, in places where few points exist the hole-filling fails.

Given several overlapping stereo reconstructions from multi-views, Hirschmüller (2008) proposes to locally merge neighbouring disparity images by the median average. Kuhn et al. (2016) presented another straight forward but efficient fusion technique. The authors remove redundancy and noise via a probabilistic aggregation step over an octree-based occupancy grid. In analogy, Wenzel et al. (2014) accumulates the result of multi stereo matching in an out-of-core octree. The points in high density zones are then filtered by investigating the imaging geometry.

Finally, there are also methods that perform fusion directly from the image data. Strecha et al. (2006) simultaneously analyse multiple overlapping images to obtain a complete and occlusion-free reconstruction. The solution is global and constrained merely by the photometric cues. Another global example was presented in Vogiatzis et al. (2005). The energy becomes a function of the photo-consistency between respective pixels, the distance function and regularizing term defined over a Markov Random Field graph. The final surface results from a discrete optimisation algorithm such as Graph-cuts. Given a visual hull Vogiatzis et al. (2005) are also capable of modelling the surface occlusions.

2.5D surface fusion. Whereas the above methods reflect the richness of existing solutions, they are mainly applied to complex 3D scenes such as small, closed objects, cluttered indoor and outdoor scenes or aerial acquisitions. Examples of local surface fusion applied to satellite images are often based on filtering and mean averaging. For instance, Reinartz et al. (2005) fuses digital surface models (DSMs) generated from SPOT-5 and SRTM by mean averaging, accompanied by the respective DSM quality indicator. Karkee

et al. (2008) take two DEMs generated within the SRTM and ASTER missions and completes them in void places with an erosion algorithm. Xu et al. (2010) create a hybrid product being a weighted average of a TerraSAR-X DEM and a DEM calculated from ALOS imagery. In all cases 2.5D surfaces generated from heterogeneous sensors are exploited to produce a more complete 3D surface.

Next, Papasaika et al. (2011) conceived a local fusion using a dictionary of implicit functions. Kuschk and dAngelo (2013) adopted a global energy formulation to fuse depth maps obtained by heterogeneous satellite sensors. The algorithm is a variational method, inspired by Pock et al. (2011). Using synthetic datasets and a World-view 2 images they showed superiority of their approach to the median-based fusion. Finally, Kuschk et al. (2016) extend their previous work to implement a new framework that is multi-resolution, allows for inclusion of quality indicators and favours locally planar surfaces.

3. Reconstruction pipeline

The reconstruction pipeline divides into three parts: geolocation refinement, multiple multi-view dense 3D reconstruction and the fusion (cf. Fig. 1).

The input geolocation parameters are first refined in a RPC-based bundle block adjustment routine, with or without the ground control points. Next, the per triplet depth maps are generated and transferred to the reference frame (RF) of the terrain geometry. Since the y-parallaxes between all images in the block had been removed, no further co-registration is necessary and the reconstructions are perfectly aligned. Finally, the fusion algorithm sweeps across the 2.5D scene to compress the data, remove the outliers, and pick the most probable Z-value for each position.

3.1. Refinement of sensor's geolocation

The physical mathematical model describing VHR satellites is represented by the central perspective projection in the across-track direction and the orthogonal projection along the track (i.e. the pushbroom sensor). The time-dependent collinearity equations describing this model are complex, therefore a common practice is to replace it with an empirical model. The standard replacement model is a set of rational polynomial functions (RPC) estimated from the physical model, with a precision close to that of the reference model (Tao and Hu, 2001).

The RPCs are nowadays always furnished with the images. Depending on the satellite's platform, they guarantee an absolute geolocation accuracy in range of a few meters (Oh and Lee, 2015), the errors being due to the inaccuracies of on-board direct georeferencing devices that measure position, attitude and the velocity of the satellite (Fraser and Hanley, 2005). Moreover, for large image blocks and multi-temporal acquisitions RPCs reveal inconsistencies between the stereo-pair, triplets, etc., i.e. their relative orientation is poorly known, and subject to some refinement.

MicMac resolves the refinement procedure similarly to the *bias compensation* method (Fraser and Hanley, 2003; Grodecki and Dial, 2003). Two polynomial corrections functions are defined in image space and estimated via a bundle block adjustment routine. The pixel displacement caused by the polynomials are constrained to be "explainable" by displacements a small sensor rotation would have caused (Rupnik et al., 2016). Tie points and ground control points are the two accepted observations groups.

3.2. Multi-view reconstruction and transfer to terrain geometry

With a set of oriented satellite images the 3D scene is inferred with a semi-global multi-view stereo reconstruction algorithm implemented in MicMac (Hirschmüller, 2008; Pierrot-Deseilligny and Paparoditis, 2006; Rupnik et al., 2017). We choose to perform the reconstruction in triplets as it is more robust than stereo processing and permits to automatically detect the occlusions. For each triplet the restitution is defined in the image coordinate system of the selected master. MicMac calculates the cost of the candidate depths in numerous ways – as a function of the correlation score calculated symmetrically between all possible pairs; calculated between pairs involving the master image only; or between all possible pairs while privileging correlation scores that involve the master. In the performed tests the first strategy was chosen, e.g. for a triplet, the cost of a depth of a given master image position depended on mean of three correlations scores. The triplets were selected favouring the B/H around 0.15.

Next, the individual reconstruction are transferred to an absolute RF defined as parallel to the ground, with the Z-coordinate pointing towards the satellites (i.e. terrain geometry). The resolution of this RF is set to a mean GSD calculated on all satellite images. The individual reconstructions are moved from image geometry to the terrain geometry by constructing a triangulated mesh in the image geometry, followed by linearly interpolating the Z-values for each position in the absolute RF.

Occlusions caused by objects or shadows and non-textured areas (e.g. water) introduce noise and outliers to the final 3D reconstruction. If an initial scene geometry is known, a technique known as *Z-buffer* can be adopted (Faugeras and Keriven, 2002) to identify them and handle accordingly. In the SGM framework, occlusions are typically predicted by performing a symmetric consistency check in a stereo couple (Hirschmüller, 2008). Otherwise,

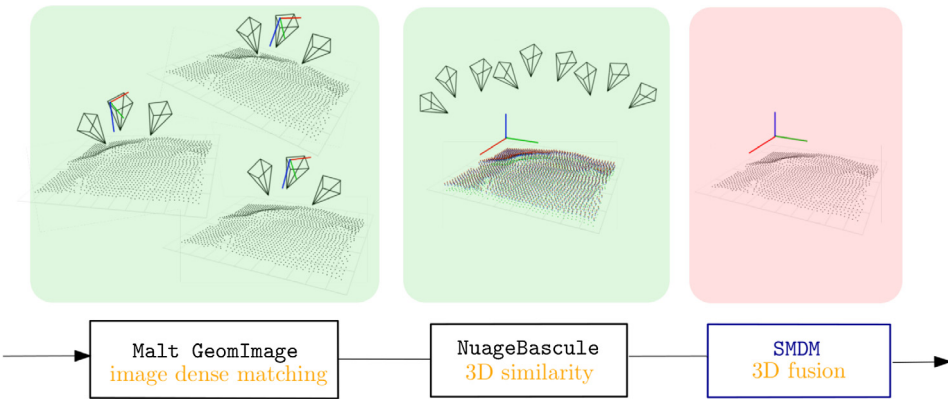


Fig. 1. The 3D reconstruction pipeline from satellite imagery in MicMac: a set of images is grouped into n-tuples and local reconstruction are carried out (`Malt GeomImage`); the reconstructions are transferred to a reference frame of the terrain geometry (`NuageBascule`) and fused (`SMDM`).

a default correlation threshold can be used to separate the correlated from non-correlated zones (Goesele et al., 2006). A more elaborate approach is to employ robust similarity statistics, for instance by exploiting multi-views and formularing the correlation score as a robust combination of each view (Vogiatzis et al., 2005).

MicMac handles occlusions (non-correlating zones in a more general term) in numerous ways. Apart from the visibility check based on the *Z-buffer* approach, or a mask defined by the user, detection of occlusions is also embedded within the SGM. Similarly to Campbell et al. (2008), an extra state e_{NC} (cf. Fig. 2 and Eq. (2)) is introduced to the multi-direction dynamic programming. The initial masks (i.e. at the lower-most level of the pyramid) contain no occlusions, and the cost of assigning the e_{NC} state as well as the cost of transition are set to elevated values. As the matching progresses to higher pyramid levels, the mask may evolve as the solver may choose the e_{NC} be the least cost state, hence, identify occlusions. Once identified, to force the solver keep the already identified masks as the least costs states, MicMac assigns them an arbitrary zero cost while the remaining states are given inversely an elevated cost.

Refer to Fig. 3 for an example representing the performance of this SGM-embedded automated detection of non-correlated areas. Note that the suppression of noise at the building edges and the constant depths within the non-textured zones.

3.3. Fusion algorithm

At this stage the algorithm has at the disposal several depth maps registered in some absolute RF, with their respective normalized cross-correlation (NCC) maps and 2D occlusion masks. The objective is to combine all this information in an optimal way and obtain a single depth map.

The processing workflow is presented in Fig. 4. To start with, the dataset is partitioned into smaller processing blocks that are in the following executed in parallel. The algorithm starts by depth clustering and outlier removal. It then identifies the most probable depth within every cluster and launches the multi-directional dynamic programming. The dynamic programming performs the typical set of operation, i.e. initialization of the cost structure; cost fill-in; cost aggregation; and lowest cost depth selection.

3.3.1. Parallelization

Prior to launching the algorithms relevant to fusion, the 3D scene is partitioned into several blocks to be further processed in parallel. It allows for a more efficient use of the processor's cores as well as avoids the memory overflow. To assure the continuity at the borders, each block is dilated by $\sim 10\%$. Within the final fused depth map the dilated zones are ignored.

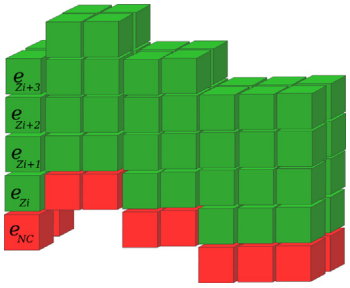


Fig. 2. The cost structure with $N+1$ states. The green states correspond to the candidate depths/disparities, the red state is the extra state reserved for non-correlated, masked-out areas. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.3.2. Depth clustering and outlier removal

The goal of the clustering and outlier removal is twofold:

- to reduce the data quantity
  (as the fusion in Section 3.3.3 has a cost of $P \cdot k \cdot N^2$ where P is the number of pixels, k number of exploited directions and N the number of depth maps);
- to robustly smooth the surface.

Each planimetric position of the 3D scene (equivalent of a pixel in a 2.5D depth map) gets an associated cell (cf. Fig. 5). A cell is composed of N ascending sorted piles, where N represents the number of depth maps. Hence, each pile stores a depth value and a corresponding weight. The weights are initialized as

$$Pds_0 = Corr^{\gamma_{corr}} \tag{1}$$

where *Corr* is the image matching NCC score transformed to the RF, γ_{corr} is a parameter that can control the relative importance of high

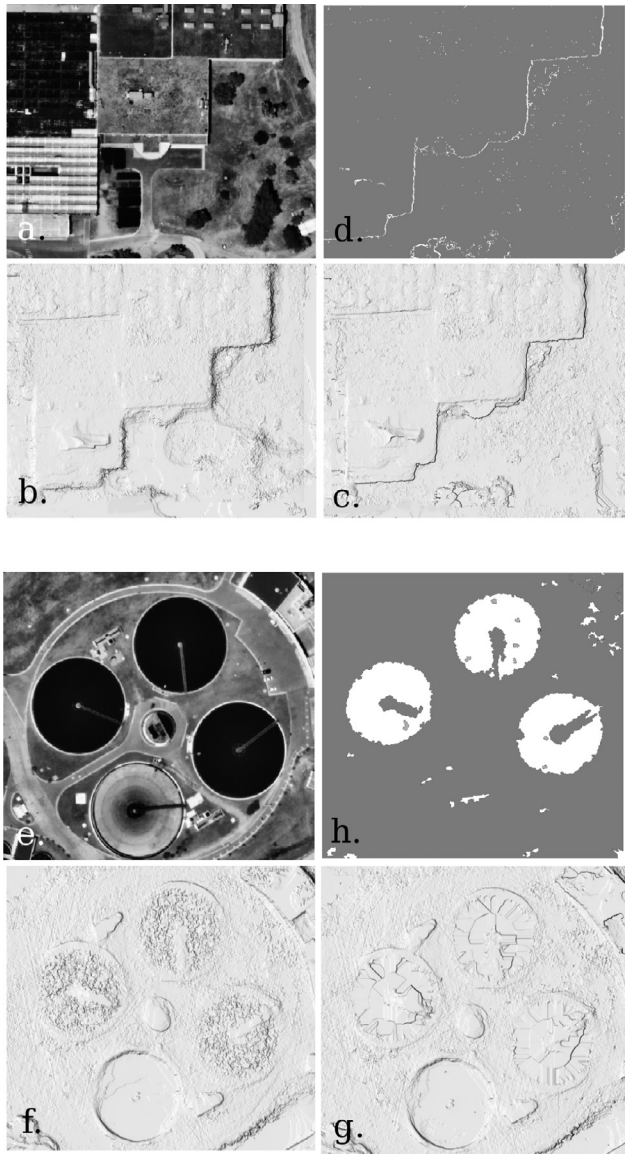


Fig. 3. Handling non-correlated zones in non-textured, water-like zone (bottom); and at building discontinuities (top). The gray shaded surface model calculated from a WV-3 triplet stereo without (b, f) and with (c, g) the e_{NC} state. (d, h) Are the SGM-inferred masks.

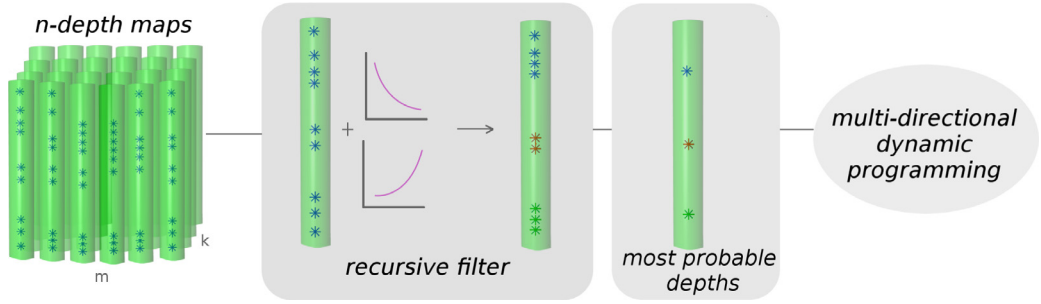


Fig. 4. The fusion algorithm workflow. For each planimetric coordinate there are n-values, corresponding to n-depth maps (here n = 9) of $m \times k$ size. All together, as many as $m \times k$ cell structures are composed. Each cell is smoothed with a recursive filter, then most probable depths are selected and passed to multi-directional dynamic programming.

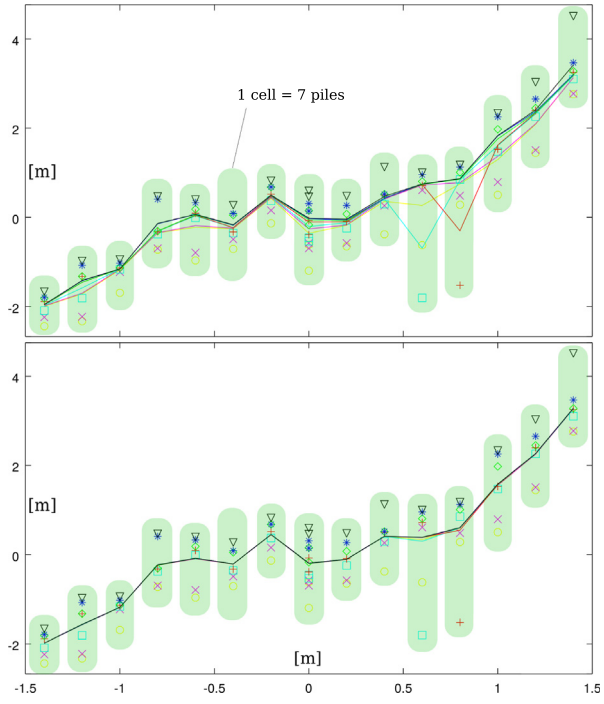


Fig. 5. Recursive exponential filtering. Vertical and horizontal axes correspond to *Z*- and *XY*-coordinates (i.e. a noisy depth map profile). For each *XY*-coordinate there is a cell with $np = 7$. A group of unique symbols describe the *Z* values furnished by the same depth map. The polylines are the filtering results, the line color is equivalent of the symbols. Top: $\sigma = 1$; bottom: $\sigma = 5$. The latter is the value used in the experiments. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

correlation scores (set to 1.0 in the performed tests). If the position in question belongs to a mask (cf. Section 3.2) the weight is set to zero.

To eliminate the outliers, the measurement noise, and limit the number of candidate depths, a fast approximation of the Gaussian filter is applied to each cell (both *Pds* and *Z*). The approximation is realized by a recursive exponential filter, see Algorithm 1. The filter is sensitive to the output GSD, and is parametrised by a σ, where the larger the σ the more conservative the filtering (cf. Fig. 5). To obtain a gaussian, the recursive filter is run two times in a row.

The end effect are smoothed clusters of depths along each cell. Finally, within each cluster a depth of the highest weight is identified and selected as the one that will be further processed.

Algorithm 1. Recursive exponential filtering by a fast gaussian approximation

np – number of piles in a cell
$Resol$ – GSD in XY
$Den = \sigma \cdot Resol \cdot \sqrt{2}$

▷ Initialization

$Pds_p^0 = Pds_0^0$
$Z_p^0 = Z_0^0$
$Pds_m^{np-1} = Pds_0^{np-1}$
$Z_m^{np-1} = Z_0^{np-1}$

▷ Propagation from "bottom to top"

for $i = 1; i < np; i++$ **do**

$Fac = (Z_0^i - Z_0^{i-1})/Den$
$Pds_p^i = Pds_0^i + Pds_p^{i-1} \cdot \exp^{Fac}$
$Z_p^i = Z_0^i + Z_p^{i-1} \cdot \exp^{Fac}$
end for

▷ Propagation from "top to bottom"

for $i = np - 2; i >= 0; i--$ **do**

$Fac = (Z_0^i - Z_0^{i+1})/Den$
$Pds_m^i = Pds_0^i + Pds_m^{i+1} \cdot \exp^{Fac}$
$Z_m^i = Z_0^i + Z_m^{i+1} \cdot \exp^{Fac}$
end for

▷ Ouput

for $i = 0; i < np; i++$ **do**

$Pds_0^i = (Pds_p^i + Pds_m^i - Pds_0^i)/np$
$Z_0^i = (Z_p^i + Z_m^i - Z_0^i)/(np \cdot Pds_0^i)$
end for

3.3.3. Multi-directional dynamic programming

The fusion algorithm uses the implementation of the multi-directional dynamic programming available in the MicMac library (Pierrot-Deseilligny et al., 2016). The optimizer solves for the most probable depths by minimizing an energy functional (Hirschmüller, 2008):

$$C(S) = \sum_{k=1}^{N} C^I(e^k_{S(k)}) + C^T(e^k_{S(k)}, e^{k+1}_{S(k+1)}) \quad (2)$$

Fig. 6. The MVS benchmark dataset. Left: trajectories of the selected acquisitions; middle: the selected area; right: a LiDAR raster over the selected area.

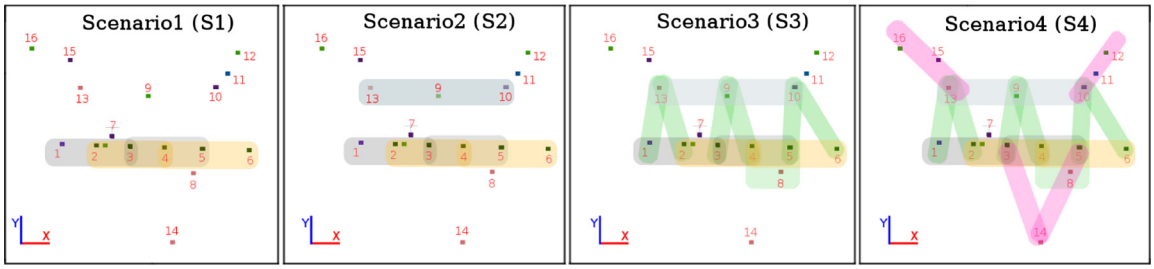


Fig. 7. Four processing scenarios of the MVS benchmark. S1: 4 triplets acquired from a single orbit (1-2-3, 2-3-4, 2-4-5, 4-5-6); S2: 5 triplets acquired from two orbits (S1 + 13-9-10); S3: 9 triplets (S2 + 1-13-2, 3-9-4, 5-10-6, 4-8-5); S4: 13 triplets (S3 + 16-15-13, 3-14-5, 10-11-12). The acquisition B/H ratios vary between 0.1 and 0.35 (e.g. 0.1, 0.2 and 0.35 for [1–2], [1–13] and [3–14], respectively).

Table 1
The MVS benchmark, vertical accuracy assessment on an **industrial zone**. μ^*, σ^* are the mean and standard deviations after removal of outliers. The outlier rejection threshold was set to 3 m. All figures are given in [m].

Scenario	μ^*	σ^*	median	NMAD	68.3% quantile	Rejection [%]
1	0.08	1.00	0.09	0.81	1.12	10.48
2	0.08	0.96	0.07	0.75	1.03	9.77
3	0.14	0.92	0.12	0.72	0.95	8.86
4	0.12	0.89	0.09	0.66	0.89	8.68

Table 2
The MVS benchmark, vertical accuracy assessment on an **residential zone**. μ^*, σ^* are the mean and standard deviations after removal of outliers. The outlier rejection threshold was set to 3 m. All figures are given in [m].

Scenario	μ^*	σ^*	median	NMAD	68.3% quantile	Rejection [%]
1	0.25	1.11	0.18	1.01	1.30	8.86
2	0.23	1.10	0.14	0.99	1.28	8.77
3	0.29	1.09	0.23	0.98	1.27	8.89
4	0.24	1.08	0.16	0.95	1.24	8.79

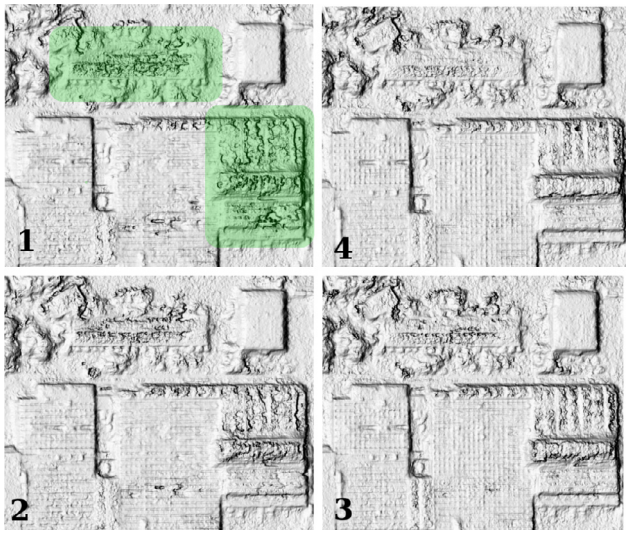


Fig. 8. *MVS benchmark*, an extract from the gray-shaded DSM over an industrial zone. Green ROI is drawn on non-textured surfaces. Scenarios (1)–(4). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where N is the number of positions (e.g. the sum of all planimetric coordinates), $e^k_{S(k)}$ corresponds to a candidate depth (i.e. a state) at position k, and $e^{k+1}_{S(k+1)}$ a candidate depth at position $k+1$, respectively. C^l is the cost of assigning a particular state to the position k (the data term), and C^T is the cost of transition between two states of the most immediate neighbours k and $k+1$ (the regularizing term). $\mathcal{C}(S)$ is the minimum cost of the optimal depths set $\mathcal{S}$, calculated across all positions.

The multi-directional aspect is carried out along any number of paths (set to 7 in the performed tests). For each depth candidate i at position l along a path of length L, the minimum cost results from an accumulation of the minimum costs along that path for positions $\{1 \dots l-1\}$, cf. Eq. (4), and the minimum costs for positions $\{L \dots l+1\}$ (in analogy to Eq. (4) but running backwards the path), subtracted by data term as it had been added twice:

$$\mathcal{C}_{min}(e^k_i) = \mathcal{C}^+_{min}(e^k_i) + \mathcal{C}^-_{min}(e^k_i) - C^l(e^k_i), \tag{3}$$

$$\mathcal{C}^+_{min}(e^{k+1}_i) = C^l(e^{k+1}_i) + Min_{j\in[1,n_k]}(\mathcal{C}^+_{min}(e^k_j) + C^T(e^k_j, e^{k+1}_i)). \tag{4}$$

To avoid the memory overflow by the permanently growing costs along the path without influencing the local minimum, the cost of each candidate depth is reduced by the minimum cost at the previous position:

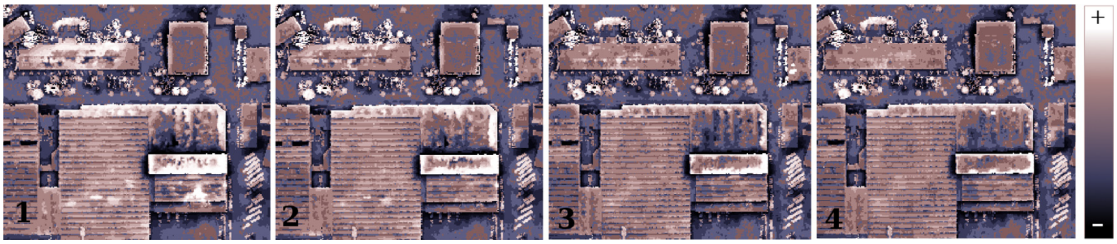


Fig. 9. *MVS benchmark*, an extract (compare Fig. 8) from the LiDAR–photogrammetric DSM difference maps over an industrial zone. Scenarios (1)–(4).

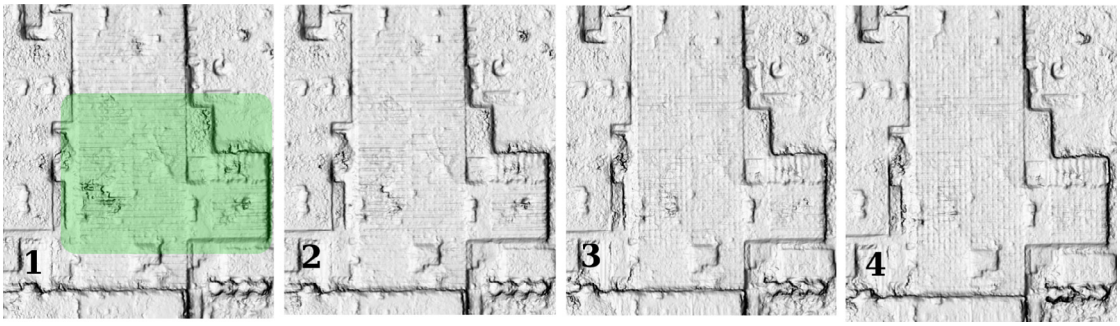


Fig. 10. *MVS benchmark*, an extract from the gray-shaded DSM over an industrial zone. Green ROI is drawn on repetitive structures. Scenarios (1)–(4). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

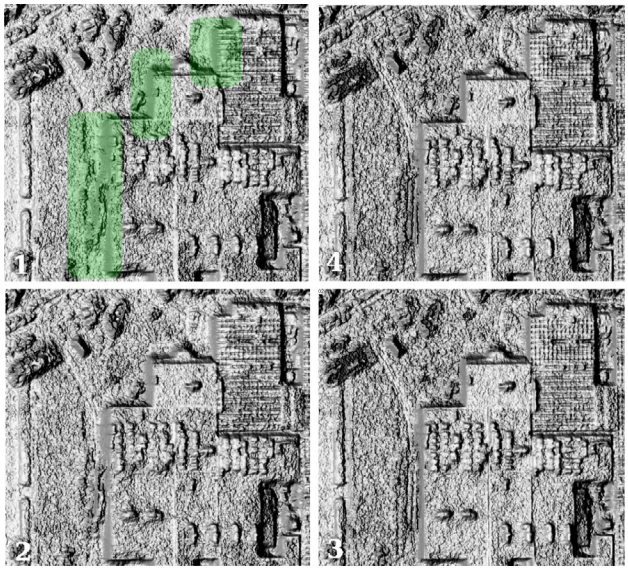


Fig. 11. *MVS benchmark*, an extract from the gray-shaded DSM over an industrial zone. Green ROI is drawn over surface discontinuities. The intensities were multiplied by a scalar to ease the visual interpretation. Scenarios (1)–(4). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$\Delta_{min}(e_i^k) = \mathcal{C}_{min}(e_i^k) - \mathcal{C}(\mathcal{S}_{min}). \tag{5}$$

Ultimately, the costs along all paths $r \in R$ are summed and the optimal depth at position k becomes the one with the smallest cost assigned, see Eq. (6). During the optimisation all costs Δ_{min} are stored in a dynamic structure, i.e. its size adapts to the number of depth candidates at each position.

$$\min_{e^k_{i\in S(k)}} \left\{ \sum_{r=1}^{r=R} \Delta_{min}(e^k_{i\in S(k)}) \right\} \tag{6}$$

The fusion cost structure is initialized with the list of cells, cleaned from noise and outliers in the previous processing steps (see Section 3.3.2 and Fig. 5). The piles within a cell provide with the possible states $S(k)$ for each k^{th} position, and their corresponding weights form the data term:

$$C^l(e^k_{S(k)}) = 1 - Pds_{e^k_{S(k)}}. \tag{7}$$

The regularising term is a concave function parametrised by the depth difference, σ and α, cf. Eq. (8). The parameters implement the following behaviour: the cost will decrease with the diminishing σ (i.e. more confidence will have more influence on the solution); on the contrary, bigger α values will increase the cost and contribute to faster growing penalties for larger dZ entries (e.g. applicable to smooth surfaces such as rural zones).

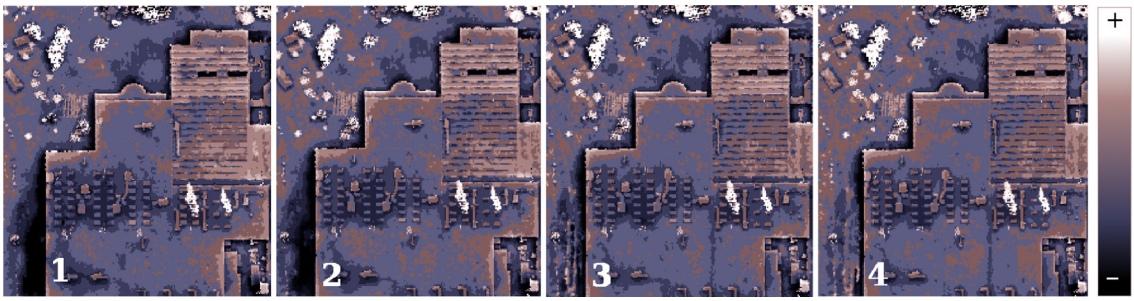


Fig. 12. *MVS benchmark*, an extract (compare Fig. 11) from the LiDAR–photogrammetric DSM difference maps over an industrial zone. Scenarios (1)–(4).

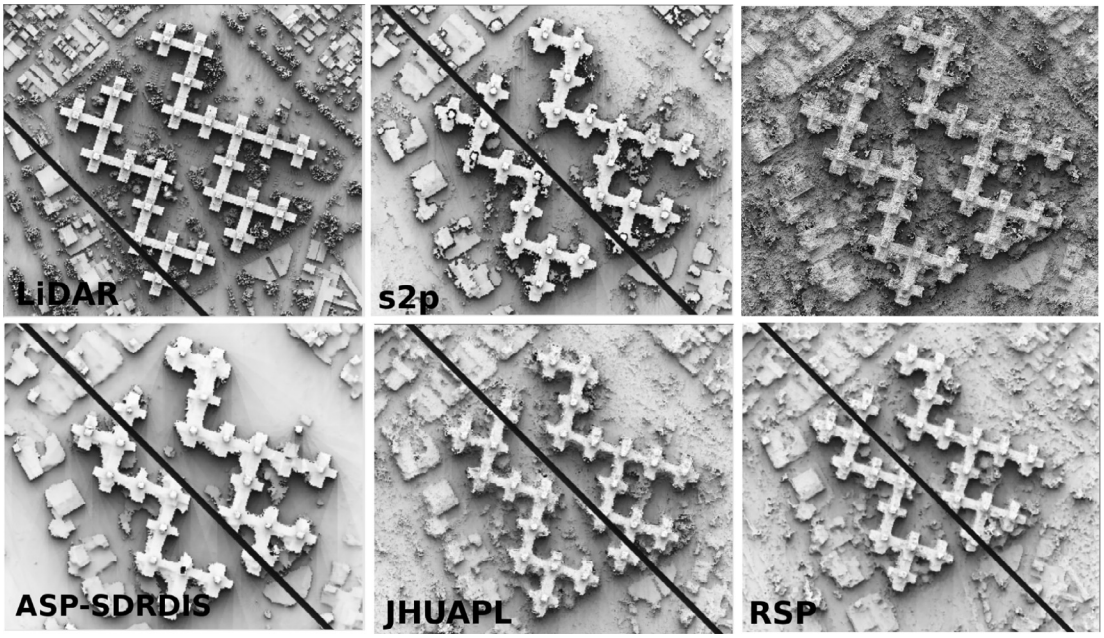


Fig. 13. IARPA's Multi-View Stereo 3D Mapping Challenge contest results presented in Bosch et al. (2017), the LiDAR ground truth and the MicMac result.

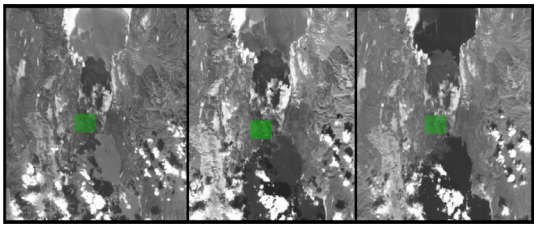


Fig. 14. *Video dataset*. First, middle and last images of the video sequence. The green ROI corresponds to the area used in the experiments. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

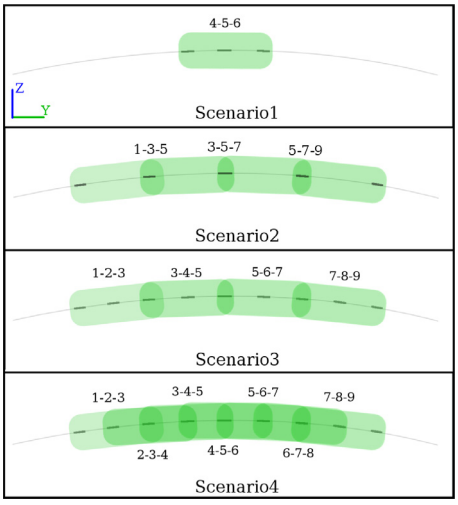


Fig. 15. Four processing scenarios of the *video dataset*. S1: a single triplet with $B/H \approx 0.1$; S2: 3 triplets, within triplet $B/H \approx 0.2$, between triplets $B/H \approx 0.2$; S3: 4 triplets, within triplets $B/H \approx 0.1$, between triplets $B/H \approx 0.2$; S4: 7 triplets, within & between triplets $B/H \approx 0.1$.

Table 3
Video dataset, vertical accuracy assessment. μ^*, σ^* are the mean and standard deviations after removal of outliers. The outlier rejection threshold was set to 3 m. All figures are given in [m].

Scenario	μ^*	σ^*	median	NMAD	68.3% quantile	Rejection [%]
1	0.34	0.53	0.35	0.52	0.57	0.25
2	0.12	0.56	0.0	0.38	0.42	2.30
3	–	–	–	–	–	–
4	0.17	0.22	0.17	0.25	0.22	0.08

$$C^T(e_j^k, e_i^{k+1}) = \left[\sqrt{1 + \frac{dZ}{\sigma_{reg}}} - 1\right] \cdot 2 \cdot \sigma_{reg} \cdot \alpha \tag{8}$$

where $dZ_{i,j}^{k,k+1} = (e_j^k - e_i^{k+1})/GSD$.

4. Results

The reported results are quantitative and qualitative. The quantitative measures correspond to the vertical accuracy calculated as a difference of LiDAR (DSM_{LiDAR}) and photogrammetric DSMs (DSM_{Pho}). The mean μ and standard deviation σ are two quality indicators adapted to errors following a normal distribution. Since photogrammetric DSMs typically contain outliers, we remove them with a 3 m-threshold and report the μ^*, σ^* on the filtered data.

Another approach considers sample quantiles to take into account non-normal distributions and resist the outliers. The median (50% quantile), Normalized Median Absolute Deviation (*NMAD*), and 68.3% quantile fall in this category (Höhle and Höhle, 2009).

As for qualitative results, gray shaded DSMs and LiDAR–photogrammetric DSM difference maps are shown.

All computations were done in MicMac. Conversion of the LiDAR point cloud to a raster format was done in LASTools (Isenburg et al., 2006).

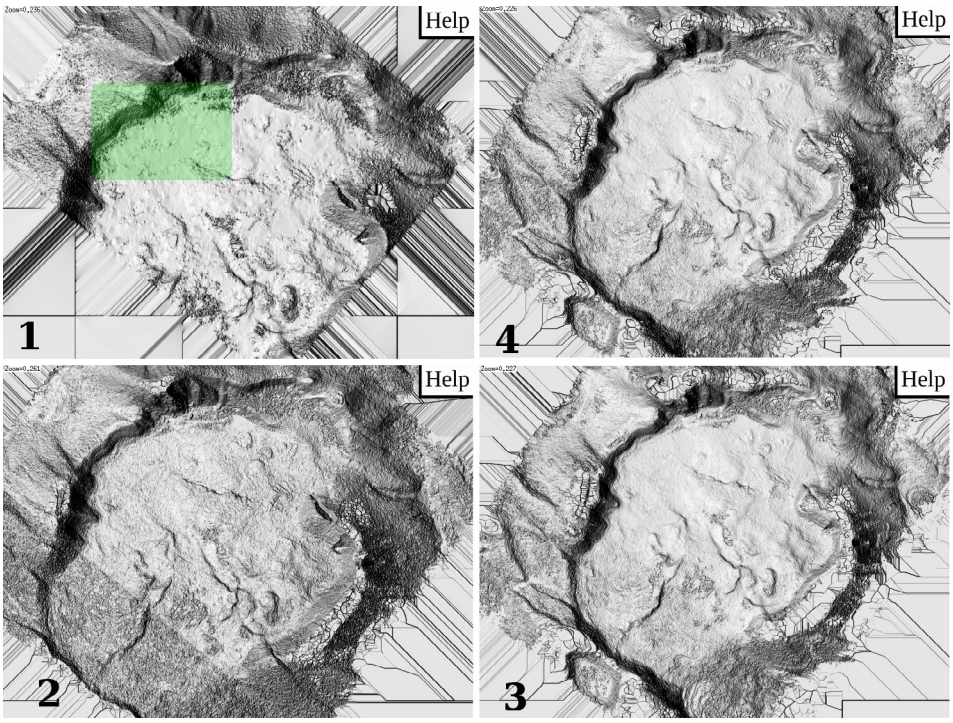


Fig. 16. *Video dataset*, the gray-shaded DSMs. Then intensities were multiplied by a scalar in order to ease the visual interpretation. The green ROI corresponds to an extract presented in Fig. 17. Scenarios (1)–(4). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

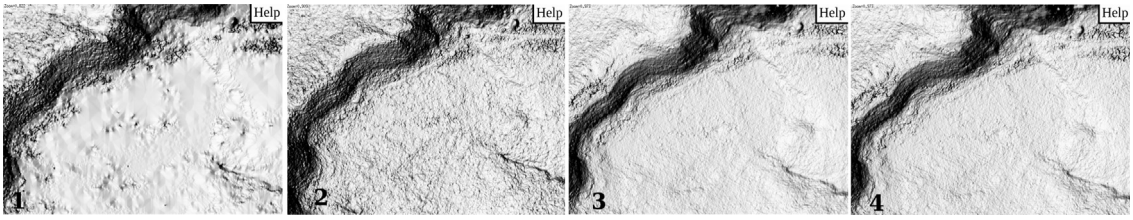


Fig. 17. *Video dataset*, an extract from the gray-shaded DSM (compare Fig. 16). Then intensities were multiplied by a scalar in order to ease the visual interpretation. Scenarios (1)–(4).

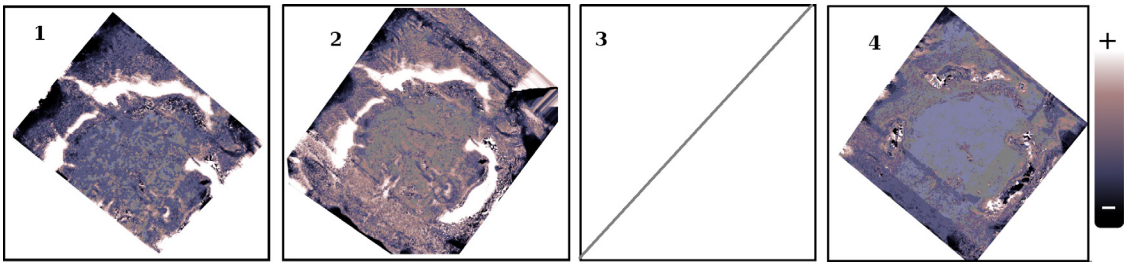


Fig. 18. *Video dataset*, DSM difference maps with the ''reference" DSM (*scenario3*).

Urban area. The urban dataset is a collection of multi-temporal, multiple view stereo panchromatic images captured with the WorldView-3 satellite (GSD = 30 cm), accompanied by a LiDAR ground truth (20 cm raster) (cf. Fig. 6). It was made available publicly within the framework of IARPA's Multi-View Stereo 3D Mapping Challenge[2] (Satellite Benchmark JHU/APL, 2016). We refer to it as the *MVS benchmark*.

The RPC parameters of the entire dataset (i.e. including images from all scenarios) were adjusted in a single bundle adjustment routine without the use of ground control points. Four acquisition scenarios were selected for the evaluation, see Fig. 7. In each scenario the multi-view reconstructions are grouped in image triplets, starting with a 4-triplets combination in a single orbit (S1) and ending with 13 triplets from multiple orbits (S4). Moreover, within all scenarios we distinguish industrial and residential area types (cf. Tables 1, 2).

The positions of the LiDAR and the photogrammetric DSM rasters are of different resolutions and do not coincide in space. The DSM_{Pho} was hence resampled to the resolution of the DSM_{LiDAR}, whereas the new positions were interpolated with a bilinear inter-

[2] www.jhuapl.edu/satellite-benchmark.html.

polator. Note the theoretical accuracy of the derived difference maps: $\sigma^2_{DSM_{LiDAR-Pho}} = \sigma^2_{DSM_{LiDAR}} + \sigma^2_{DSM_{Pho}}$, thus if we assume $\sigma_{DSM_{Pho}} = 0.75$ m and $\sigma_{DSM_{LiDAR}} = 0.2$ m, then $\sigma_{DSM_{LiDAR-Pho}} \approx \sigma_{DSM_{Pho}}$.

Quantitative results of the fusion on the *MVS benchmark* in an industrial zone are given in Table 1. The systematic shift between DSM_{LiDAR} and DSM_{Pho} encoded in the means and medians revolves around 1/3 GSD. The standard deviations and *NMADs* range from 1.0 m ($\sigma^*, SC1$) to 0.66 m (*NMADs*, *SC*4), with the tendency to decrease when more triplets are merged. The same tendency applies to the amount of the detected and removed outliers. The magnitude of *NMADs* is generally smaller than that of the standard deviations suggesting that the underlying error distribution does not follow a gaussian curve. As far as the residential zone is concerned, similar trends manifest with slightly elevated quality indicators as a consequence of the smaller but denser housing (cf. Table 2).

Visual assessment of the gray shaded DSM_{Pho} and the difference maps reveal details not captured by the quantitative measures. In Figs. 8 and 9 it is observed that employing additional triplets improves reconstruction on non-textured surfaces. The performance on repetitive structures is also excelled as is visible in Fig. 10, where merely *SC*3 and *SC*4 manage to restitute the checkerboard nature of the rooftop. Finally, the results where multiple orbit acquisitions are fused prove superior on surface discontinuities by mitigating the smoothing effects on building edges as seen in Figs. 9, 11, 12.

In Fig. 13 additional comparison against the IARPA's Multi-View Stereo 3D Mapping Challenge contest results are shown. Following the evaluation methodology in Bosch et al. (2017), the metrics obtained are: $[0.66, 0.90, -0.77]$ m in offsets on XYZ, 0.82 m as the horizontal RMSE, 4.12 as the $error_{rms}$, 0.91 as the $error_{median}$ and 0.53 in the completeness score. Visual inspection reveals a noisier but apparently more detailed surface reconstruction by our algorithms.

Rural area. The rural dataset is a Pléiades video acquisition (cf. Fig. 14) comprising of 18 images, and we refer to it as the *video dataset*. The RPC parameters of the 18 images were adjusted in a single bundle adjustment routine with the use of a few ground control points. Again, four acquisition scenarios were chosen (cf. Fig. 15) with the objective to identify the optimal results in terms of accuracy, by comparing them with a reference dataset. Since no ground truth was available, out of four scenarios we picked as reference the DSM of best visual quality (*scenario3*, cf. Fig. 17).

Quantitative results are presented in Table 3. The figures are relative and communicate that with the growing number of triplets the noise is diminished. Another message is that larger B/H within the triplets (*scenario2*) contribute to higher percent of outliers as well as a non-gaussian error distribution.

More interesting is to visually inspect the outcome DSMs. It is clear from Figs. 16 and 17 that the fused products can handle surfaces of little or no texture (see the presence of smoothing artefacts in Fig. 17, *scenario1*). Adding more views completes the reconstructions but generates noise for too large B/H ratios, see *scenario2*. The same behaviour can be decoded from the difference maps in Fig. 18: flat patches in *scenario1* characteristic of the smoothing artefacts; noise and lack of detail in *scenario2*. *Scenarios3* and *scenario4* with smaller B/H, especially within the triplets, are indisputably the most optimal results.

5. Conclusions

This publication presents a 3D reconstruction pipeline adapted to modern multi-view satellite acquisitions. The pipeline follows a refinement of the sensor's geolocation (via bundle block adjustment), per n-tuplet image dense matching, the terrain geometry transfer phase and finally the fusion of individual depth maps. The fusion – the essence of this work – is formulated in a semi-global optimization framework, takes into account the occlusions and other non-correlated zones as well as quality indicators resulting from n-tuplet image dense matching. The algorithm makes no *a priori* on the resolution of the to-be-fused depth maps hence handles DSMs resulting from different resolution sensors. All the above is implemented in the free open-source software for photogrammetry – MicMac.

The pipeline was tested on two datasets acquired with Pléiades and WordView-3 satellites, and evaluated in a quantitative and qualitative manner. A slight improvement in the accuracy of the fused DSMs is observed as the number of merged views increase. The quantitative assessment, nonetheless, is incapable of revealing the true advantages of the approach. Qualitative results in form of gray shaded DSMs show that non-textured areas, and surface discontinuities are much better resolved using the fusion approach.

The tested acquisition scenarios also exposed the importance of the B/H ratios within the individual n-tuplets and between the neighbouring reconstructions. It was observed that for the former, ratios revolving around 0.1 provide best outcomes. On the other hand, the quality of results is less sensitive to the ratios between reconstructions hence with values relaxed to 0.2 optimal performance was achieved.

Future research will concentrate on the automated selection of image n-tuplets used in the image dense matching as well as their combination in the fusion process.

Acknowledgements

This research was funded by CNES via the TOSCA programme. The Pléiades video acquisition was obtained through the CNES' ISIS framework. The satellite imagery in the MVS benchmark dataset was provided by the courtesy of DigitalGlobe. The numerical computations were performed on the SCAPAD cluster processing facility at the Institute de Physique du Globe de Paris. We would like to kindly thank Myron Brown and Bosch Ruiz for providing us with the Multi-View Stereo 3D Mapping Challenge contest results.

References

Bosch, M., Leichtman, A., Chilcott, D., Goldberg, H., Brown, M., 2017. Metric evaluation pipeline for 3d modeling of urban scenes. Int. Arch. Photogramm. Remote Sens. Spatial Inform. Sci. 42, 239.

Campbell, N.D., Vogiatzis, G., Hernández, C., Cipolla, R., 2008. Using multiple hypotheses to improve depth-maps for multi-view stereo. In: European Conference on Computer Vision. Springer, pp. 766–779.

Curless, B., Levoy, M., 1996. A volumetric method for building complex models from range images. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques. ACM, pp. 303–312.

Faugeras, O., Keriven, R., 2002. Variational principles, surface evolution, PDE's, level set methods and the stereo problem. IEEE.

Fraser, C., Hanley, H., 2003. Bias compensation in rational functions for IKONOS satellite imagery. Photogramm. Eng. Remote Sens. 69 (1), 53–57.

Fraser, C., Hanley, H., 2005. Bias-compensated rpcs for sensor orientation of high-resolution satellite imagery. Photogramm. Eng. Remote Sens. 71 (8), 909–915.

Fuhrmann, S., Goesele, M., 2011. Fusion of depth maps with multiple scales. ACM Transactions on Graphics (TOG), vol. 30. ACM, p. 148.

Goesele, M., Curless, B., Seitz, S.M., 2006. Multi-view stereo revisited. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2. IEEE, pp. 2402–2409.

Grodecki, J., Dial, G., 2003. Block adjustment of high resolution satellite images described by rational polynomials. Photogramm. Eng. Remote Sens. 69 (1), 59–68.

Hirschmüller, H., 2008. Stereo processing by semiglobal matching and mutual information. IEEE Trans. Pattern Anal. Mach. Intell. 30 (2), 328–341.

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.

Isenburg, M., Liu, Y., Shewchuk, J., Snoeyink, J., Thirion, T., 2006. Generating raster DEM from mass points via TIN streaming. Geogr. Inform. Sci., 186–198

Karkee, M., Steward, B.L., Aziz, S.A., 2008. Improving quality of public domain digital elevation models through data fusion. Biosyst. Eng. 101 (3), 293–305.

Kazhdan, M., Hoppe, H., 2013. Screened poisson surface reconstruction. ACM Trans. Graph. (TOG) 32 (3), 29.
Kuhn, A., Hirschmüller, H., Mayer, H., 2013. Multi-resolution range data fusion for multi-view stereo reconstruction. In: German Conference on Pattern Recognition. Springer, pp. 41–50.
Kuhn, A., Huang, H., Drauschke, M., Mayer, H., 2016. Fast probabilistic fusion of 3d point clouds via occupancy grids for scene classification. In: XXIII ISPRS Congress, Technical Commission III, vol. 3, pp. 325–332.
Kuschk, G., dAngelo, P., 2013. Fusion of multi-resolution digital surface models. ISPRS-Int. Arch. Photogramm. Remote Sens. Spatial Informat. Sci. 1 (3), 247–251.
Kuschk, G., d'Angelo, P., Gaudrie, D., Reinartz, P., Cremers, D., 2016. Spatially regularized fusion of multiresolution digital surface models. IEEE Trans. Geosci. Remote Sens. 55 (3), 1477–1488.
Oh, J., Lee, C., 2015. Automated bias-compensation of rational polynomial coefficients of high resolution satellite imagery based on topographic maps. ISPRS J. Photogramm. Remote Sens. 100, 14–22.
Papasaika, H., Kokiopoulou, E., Baltsavias, E., Schindler, K., Kressner, D., 2011. Fusion of digital elevation models using sparse representations. In: Photogrammetric Image Analysis. Springer, pp. 171–184.
Pierrot-Deseilligny, M., Paparoditis, N., 2006. A multiresolution and optimization-based image matching approach: an application to surface reconstruction from spot5-hrs stereo imagery. In: ISPRS Workshop On Topographic Mapping From Space, vol. 36(1). Ankara, Turkey.
Pierrot-Deseilligny, M., Rupnik, E., Girod, L., Belvaux, J., Maillet, G., Deveau, M., Choqueux, G., 2016. Micmac, apero, pastis and other beverages in a nutshell.
Pock, T., Zebedin, L., Bischof, H., 2011. TGV-fusion. In: Rainbow of computer science. Springer, pp. 245–258.
Reinartz, P., Müller, R., Hoja, D., Lehner, M., Schroeder, M., 2005. Comparison and fusion of DEM derived from SPOT-5 HRS and SRTM data and estimation of forest heights. In: Proc. EARSeL Workshop on 3D-Remote Sensing, Porto, vol. 1.
Rupnik, E., Daakir, M., Pierrot Deseilligny, M., 2017. Micmac – a free, open-source solution for photogrammetry. Open Geospatial Data Softw. Stand. 2 (14), 1–9.
Rupnik, E., Pierrot-Deseilligny, M., Delorme, A., Klinger, Y., 2016. Refined satellite image orientation in the free open-source photogrammetric tools APERO/MICMAC. ISPRS Ann. Photogramm. Remote Sens. Spatial Inform. Sci.
Satellite Benchmark JHU/APL, 2016. A multiple view stereo benchmark for satellite imagery.
Schindler, K., Papasaika-Hanusch, H., SCHüTZ, S., Baltsavias, E., 2011. Improving wide-area DEMS through data fusion–chances and limits. In: Proceedings of the Photogrammetric Week, vol. 11. pp. 159–170.
Strecha, C., Fransens, R., Van Gool, L., 2006. Combined depth and outlier estimation in multi-view stereo. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2. IEEE, pp. 2394–2401.
Tao, C.V., Hu, Y., 2001. A comprehensive study of the rational function model for photogrammetric processing. Photogramm. Eng. Remote Sens. 67 (12), 1347–1358.
Ummenhofer, B., Brox, T., 2015. Global, dense multiscale reconstruction for a billion points. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1341–1349.
Vogiatzis, G., Torr, P.H., Cipolla, R., 2005. Multi-view stereo via volumetric graph-cuts. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, vol. 2. IEEE, pp. 391–398.
Vu, H.-H., Labatut, P., Pons, J.-P., Keriven, R., 2012. High accuracy and visibility-consistent dense multiview stereo. IEEE Trans. Pattern Anal. Mach. Intell. 34 (5), 889–901.
Wenzel, K., Haala, N., Fritsch, D., 2014. Stereo model selection and point cloud filtering using an out-of-core octree. Int. Arch. Photogramm. Remote Sens. Spatial Inform. Sci. 40 (3), 373.
Xu, C., Wei, M., Griffiths, S., Mercer, B., Abdoullaev, R., 2010. Hybrid DEM generation and evaluation from spaceborne radargrammetric and optical stereoscopic DEMS. In: Proc. of Canadian Geomatics Conference.
Zach, C., 2008. Fast and high quality fusion of depth maps. In: Proceedings of the International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT), vol. 1.
Zach, C., Pock, T., Bischof, H., 2007. A globally optimal algorithm for robust TV-l 1 range image integration. In: IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007. IEEE, pp. 1–8.