Automatic 3D Reconstruction from Multi-Date Satellite Images

Gabriele Facciolo Carlo de Franchis Enric Meinhardt-Llopis

École Normale Supérieure Paris-Saclay

http://gfacciol.github.io/multi-date-stereo

Abstract

We propose an algorithm for computing a 3D model from several satellite images of the same site. The method works even if the images were taken at different dates with important lighting and vegetation differences. We show that with a large number of input images the resulting 3D models can be as accurate as those obtained from a single same-date stereo pair. To deal with seasonal vegetation changes, we propose a strategy that accounts for the multi-modal nature of 3D models computed from multi-date images. Our method uses a local affine camera approximation and thus focuses on the 3D reconstruction of small areas. This is a common setup in urgent cartography for emergency management, for which abundant multi-date imagery can be immediately available to build a reference 3D model. A preliminary implementation of this method was used to win the IARPA Multi-View Stereo 3D Mapping Challenge 2016. *Experiments on the challenge dataset are used to substantiate our claims.*

1. Introduction

The number of optical Earth observation satellites has increased drastically over the past decade[1], driven by the need to monitor changes on the surface of the Earth. As a result, the amount of acquired images has grown to the point that nowadays many sites (usually urban areas) are captured several times per year. However, most of these images are taken at different dates and thus are not intended for computation of 3D models. But monitoring the Earth's surface is a three-dimensional problem and 3D models have a variety of applications such as ortho-rectification of images or support cartography for emergency management. The goal of this paper is to present an algorithm to exploit such large archives of single-date images to compute the best possible 3D model with reasonable computational cost.

[1]For example, Pléiades, Landsat 8, Worldview 3, Sentinel-2, and many more launched by private companies such as Planet, AstroDigital, Urthecast, BlackSky, Hera Systems, and Satellogic.

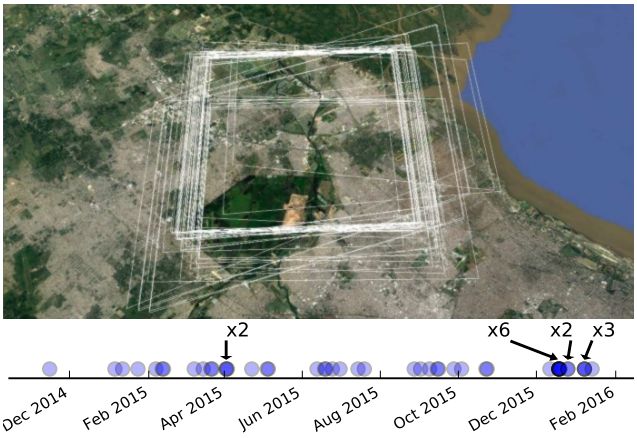


Figure 1. Footprints and dates of the 47 images of the IARPA challenge dataset [2]. The images cover the North part of Buenos Aires and were acquired over a period of 14 months. Only four groups of images were taken during the same orbit: two pairs, one triplet, one hextuple.

Approaches for 3D reconstruction from multiple views can be grouped in two classes. On the one hand, *true multiview* methods tackle the multiview triangulation problem for all images simultaneously [26, 12, 24]. On the other hand, *multiview stereo* methods use binocular stereo to process several image pairs independently and then fuse the resulting 3D models [19, 21]. It was already observed [23] that this second strategy may give better results than sophisticated true multiview methods [26]. To correct inaccuracies in the camera models, all of these methods rely on bundle adjustment [27, 30, 13] which in turn relies on detecting a sufficient quantity of accurate inter-image tie-points. This can be an issue with multi-date images, especially when restricted to small regions of interest.

In this paper, we argue in favor of *multiview stereo without bundle adjustment*: we compute independent 3D models from pairs of images with binocular stereo, without any prior bundle adjustment. It is then easy to align and fuse the multiple 3D models. This is possible thanks to a local affine camera approximation [11, 13, 22, 8, 28] implying that on small image regions, the 3D models differ by a 3D

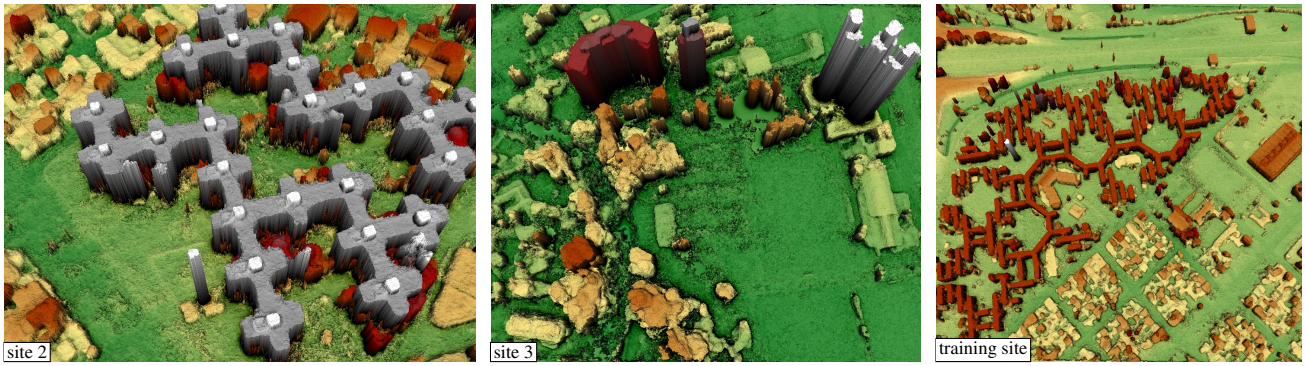


Figure 2. 3D models obtained from 47 Worldview-3 images by fusion of 50 well-chosen stereo pairs.

translation only.

We show that a 3D model computed from multi-date images by pairwise fusion can be as accurate as a 3D model computed from a same-date stereo pair, and study the best way to select image pairs and fuse the resulting models.

We highlight the importance of vegetation by showing how seasonal changes among the images affect the quality of the final reconstruction, and propose a fusion strategy that accounts for the multi-modal nature of 3D models computed from images taken at different dates.

1.1. IARPA MVS Challenge Dataset and Evaluation

This work was motivated by the release of a public benchmark dataset for multiple view stereo mapping using multi-date satellite images [2]. This dataset, which supported the *IARPA Multi-View Stereo 3D Mapping Challenge*, includes 47 DigitalGlobe WorldView-3 panchromatic images of a 100 square kilometer area near San Fernando, Argentina (see Figure 1). The images have a 30 cm nadir resolution and were acquired over a period of 14 months. Most of the images were taken at different dates. Nearly all the images are clear sky. However, the quality is not consistent: the winter images are considerably noisier, and the images with large incidence angles suffer from a loss of resolution in the range direction.

The dataset also includes 20 cm resolution airborne lidar ground truth for a 20 square kilometer subset of the covered area. It comes with a program for computing the *completeness* and *accuracy* of any 3D model, by comparing it to the lidar ground truth. Completeness is defined as the percentage of lidar points wose error is less than 1 meter, and accuracy is the root mean square error of all the computed points. Since completeness implies a certain accuracy (below 1 meter) for a set of pixels, it comes to no surprise that both metrics exhibit a strong (negative) correlation (see Section 2.1). For this reason, this paper uses completeness as the main quality measure. Figure 2 shows some results of the method proposed in this paper over the IARPA dataset.

1.2. RPC Camera Model and Pointing Error

Each satellite image is provided with a Rational Polynomial Coefficients (RPC) camera model [9], and other metadata such as the exact acquisition date or the direction of the sun. The RPC model combines the intrinsic and extrinsic parameters of the pushbroom system in a pair of rational polynomial functions that approximate the mapping from 3D space points given as (latitude, longitude, height) to 2D image pixels: $P_n : \mathbf{R}^3 \to \mathbf{R}^2$ (named *projection*), and its inverse: $L_n : \mathbf{R}^2 \times \mathbf{R} \to \mathbf{R}^3$ (*localization*). Both rational functions have degree 3 (for a total of 160 coefficients per image). The RPC model approximation has sub-millimetric accuracy for scenes of size up to 20 km $\times$ 20 km [9].

The RPC functions allow to triangulate the position of a 3D point that has been identified on two images. If the point (i, j) of image n corresponds to the point (i', j') of image n', then for some height h we have $L_{n'}(i', j', h) = L_n(i, j, h)$, or equivalently

$$(i', j') = P_{n'}(L_n(i, j, h)). \tag{1}$$

By solving equation (1) for h we find the height of the 3D point, and hence its 3D position.

Although the RPC are accurate, the model they encode is subject to measurement errors (mainly for the satellite attitude angles), which translate into geopositioning errors of the triangulated points. These pointing errors can be of the order of tens of pixels in the image domain. In [13] it was shown that since the satellite camera is far from the scene (typically 700 km), the rays for individual pixels are almost parallel. Thus geopositioning errors can be corrected by applying a *bias correction offset* (i.e. a translation) for scenes of size up to 50 km $\times$ 50 km.

1.3. Related Work

Fusing DSMs (Digital Surface Models) computed independently from pairs of multi-date images was considered in [23] and was compared to a *true multiview* volumetric method [26]. The conclusion was that fusion generates better quality DSMs, i.e. with more pixels within 1 meter of the

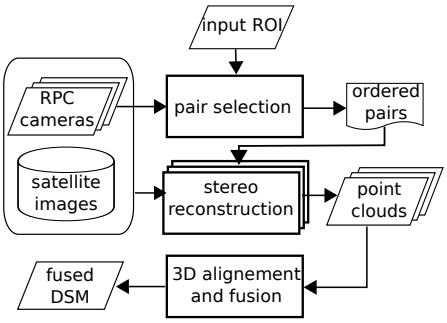
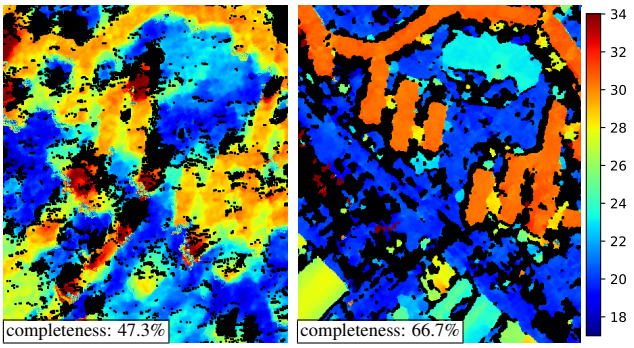


Figure 3. Flow diagram of the multi-date DSM generation pipeline used in this paper. Rational Polynomial Coefficients (RPC) camera models are used to determine the pairs to be processed. For each pair a point cloud is computed. These are then aligned and fused (see Section 2.3).



same-date (max incidence 45°) multi-date (max incidence 15°)

Figure 4. DSMs obtained from a same-date pair with large incidence, and from a multi-date pair with small incidence.

ground truth. The method proposed in [23] starts by correcting the camera models [13] of the entire image collection. This process relies on bundle adjustment and thus in detecting a sufficient quantity of inter-image tie-points, which may be hard to achieve in a multi-date setting. The method proposed in the present paper does not use bundle adjustment. Instead, it relies on a binocular stereo pipeline [6] that internally corrects the effects of relative pointing error for each pair (using pairwise image tie-points), thus producing biased DSMs. These biases are later corrected by our DSM registration step, without relying on image tie-points. Moreover, in this paper we highlight the impact of selecting and fusing few good pairs rather than computing the median of all the possible pairs as [23] does. Our method is based on a principled pair selection criterion and a fusion strategy that accounts for the multi-valued nature of a multi-date DSM.

A *multiview stereo* approach is used in [4, 19, 14] to reconstruct (and compare [14]) large scale models from sets of same-date pairs. The methods rely on bundle adjustment to align the DSMs. The fusion of the DSMs is performed by median filtering. In [28] an additional 3D model registration is applied before fusion, in order to improve the accuracy of the initial bundle adjustment. This paper also remarks that a simple 3D translation is almost always sufficient to correct WorldView-1 or WorldView-2 DSM products.

In the context of planetary science, the work of [1] gives recommendations for identifying suitable stereo pairs from a heterogeneous collection of images. Our method uses similar selection rules for the case of Earth observation satellites.

2. Proposed Multiview Reconstruction Method

Our method works by aggregating point clouds computed independently from well-chosen image pairs. A similar strategy, aggregating the DSMs obtained from all possible image pairs, was proposed in [23]. Here, we propose a new incremental method that selects and aggregates only a small fraction of all the pairs. Our method consists of three stages (Figure 3):

1. **Pair selection.** We propose a heuristic for sorting all the possible image pairs so that the first pairs of the list yield results with higher completeness measure.

2. **Stereo matching.** For each selected image pair a 3D point cloud is computed by stereo matching and triangulation. Note that each point cloud is computed independently with no need for bundle adjustment.

3. **Alignment and fusion.** The triangulated point cloud computed from each selected image pair is projected into a geographic grid and registered with the others. The small size of the reconstructed region allows to correct the geolocation errors with a simple 3D translation [13, 23, 28]. The aligned DSMs are then merged. Since the DSMs may correspond to different dates, we must assume that they are multimodal (see Section 2.3). The proposed fusion strategy accounts for this by favoring the elevation modes closer to the ground.

2.1. Selection of Image Pairs

The quality of 3D models obtained from different pairs of multi-date images varies widely. For example, Figure 4 compares the DSM from a multi-date near-nadir pair, with the DSM obtained from a same-date slanted pair. The second DSM is notably worse. Factors such as the geometric configuration of the satellites (i.e. baseline and incidence), image noise, seasonal changes, illumination, and shadows can affect the quality of the output for a given image pair. So, given a set of multi-date images, we want a criterion for sorting all the pairs according to their quality (defined by the completeness measure), and process only the first elements of this list.

To learn which factors are the most relevant for this task we computed the DSMs of the training site for all the possible image pairs (Section 2.2), and evaluated them using the ground truth data by computing completeness and accuracy. Figure 5 illustrates these quantities as cells in a matrix where rows and columns correspond to image indexes in chronological order. Note how pairs of images that are temporally close (close to the diagonal) lead to better results. Since accuracy and completeness are closely related we focus on the latter only for our evaluation.

In order to identify descriptors that can predict the completeness, we built a correlation matrix (Figure 6) between the measures and some descritptors computed from the RPC models of the images. The three most relevant are: angle between the views, maximum incidence angle, and time difference between the two images. To understand how these variables affect the completeness we partitioned this 3-parameter space and computed the average completeness for each cell as shown in Figure 7. We observed that:

1. **Temporal proximity.** Images acquired at nearby dates are more likely to yield good results. To our surprise, we also observed that images from the same season of different years also yield good results (see Figure 5).

2. **Maximum incidence angle.** When one of the two images has an incidence angle larger than 40 degrees its lower resolution degrades the result.

3. **Angle between the views.** The best results are obtained with pairs forming an angle of about 20 degrees. Angles below 5 degrees and above 45 tend to be less useful.

Based on these observations we propose a simple heuristic for sorting the image pairs. We prioritize the pairs forming angles from 5 to 45 degrees, with maximum incidence angle below 40 degrees. Within this set we sort all the pairs by increasing acquisition date difference. The remaining pairs are also sorted by increasing time difference and appended to the list.

2.2. Stereo From an Image Pair

The stereo matching was performed with an open source pipeline called S2P (Satellite Stereo Pipeline) [6]. This pipeline computes 3D point clouds from pairs of satellite images. Similarly to other open source stereo pipelines such as ASP [21] and MicMac [25], and to other works [31, 4, 5, 19], S2P is fully automatic.

But unlike most of these pipelines S2P is script-based and modular. This makes it easy to recover intermediate results and change parts. The stereo matching algorithm can for example be replaced by any other method, while S2P provides the end-to-end plumbing for tile-wise processing, camera modeling, and raster DSM synthesis. The program

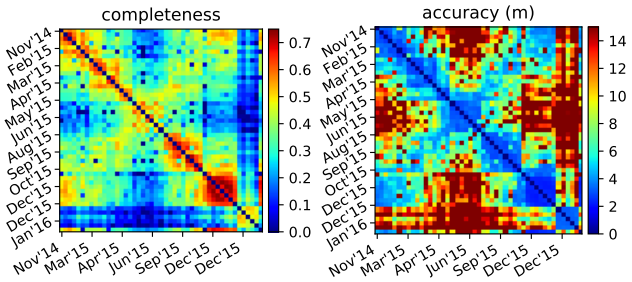


Figure 5. Completeness (percentage of pixels with error below 1 m) and accuracy (RMSE in m) of the training site reconstructions from all the possible image pairs (47×46) in the IARPA challenge dataset [2]. Rows and columns correspond to image indexes sorted by acquisition date. Note that both measures are strongly correlated (negatively).

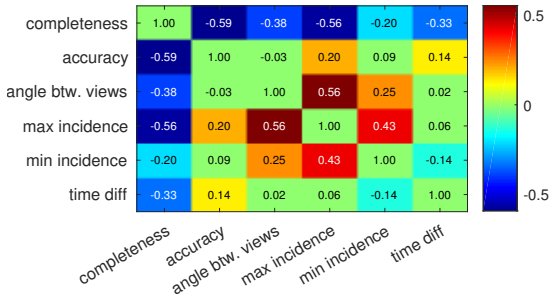


Figure 6. Pearson's correlation matrix from 2162 results on the training site. The angle between the views, maximum incidence angle, and time difference between the acquisitions are strongly correlated with the completeness.

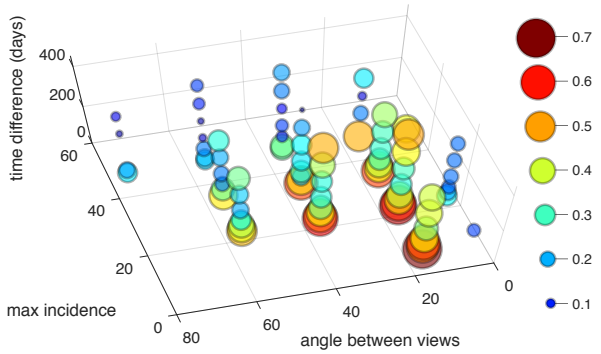
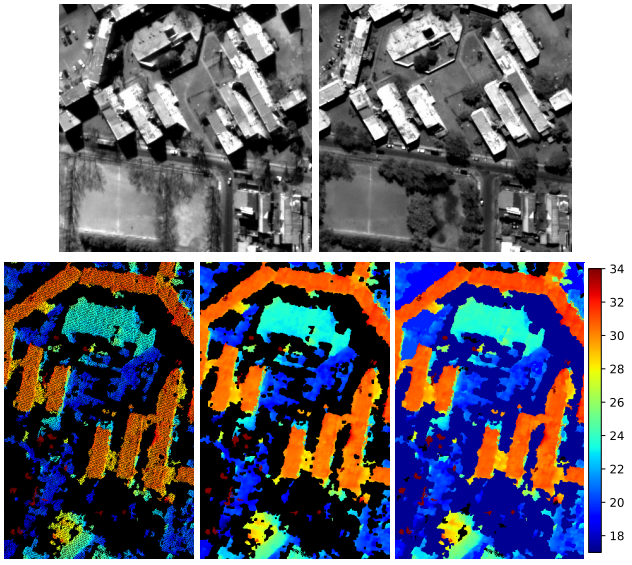


Figure 7. Average completeness (represented by color and size of the blobs) of an image pair as function of: the angle between the views, the maximum incidence of the two images, and the time difference. The averages are computed using all the possible image pairs.

works by cutting the input images into small tiles, where the camera can be assumed to be affine. Then, each pair of tiles is stereo-rectified [18] and fed to the stereo matching algorithm. Finally, the point cloud is obtained by triangulation of the stereo correspondences using the provided RPCs.



projection ($\mathbf{x}$) closing ($\mathbf{D}_p$) interpolation ($\mathbf{E}_p$)

Figure 8. Detail from a multi-date pair, taken from a similar viewpoint but in different seasons. Notice the changes in the trees and shadows. In the second row: DSM processing before fusion (steps 2-4 of Algorithm 1) for the above pair.

Stereo from multi-date pairs must deal with strong appearance changes. Although S2P is designed for same-date pairs the results on multi-date pairs can have comparable quality, provided the pair is well-chosen (see Figure 8). It has been observed [3] that the census transform [32] is robust to lighting changes and even to small rectification errors [17]. The matching algorithm [10] (a variant of SGM [16]) included in S2P uses a census-based method. Sub-pixel accuracy is achieved by sampling the disparity space with 0.5 pixel steps, and further refined by V-fit interpolation [15] of the costs around the minimum. Only consistent disparities passing the left-right check are kept.

In order to deal with the increased number of mismatches in the multi-date stereo setting we added a filtering step that removes connected disparity components smaller than 5×5 pixels. We observed that this eliminates most of the mismatch artifacts introduced by the matching algorithm [10].

2.3. Alignment and Fusion

The 3D point clouds, independently computed in the previous step, are projected, aligned, and fused in this step. The process is scalable as new pairs can be incorporated, only requiring to refresh the fusion.

Projection of point clouds into a geographic grid. We project the 3D point clouds on a geographic grid with a resolution similar to the satellite nadir GSD (ground sampling distance), which is 30 cm for the WordView 3 images provided by [2]. The algorithm computes the position of the 3D points on the geographic grid by nearest-neighbor interpolation, and stores the maximum altitude in each cell.

Algorithm 1: Alignment and fusion algorithm

Input : point clouds $\{\mathbf{C}_p\}_{p=1...P}$
Input : reference cloud index: ref
Input : geographic Region Of Interest: ROI
Output: Fused DSM

// Generate dense DSMs
for $p \in \{1, \dots, P\}$ **do**
  $\mathbf{x} \leftarrow$ PROJECTPOINTCLOUD$(\mathbf{C}_p, ROI)$
  $\mathbf{D}_p \leftarrow$ IMCLOSE$(\mathbf{x}, 3)$
  $\mathbf{E}_p \leftarrow$ INTERP5PC$(\mathbf{D}_p)$
// Align with D_{ref}
for $p \in \{1, \dots, P\}$ **do**
  $dx, dy \leftarrow \arg\max_{dx,dy}$ FNCC$(\mathbf{E}_p, \mathbf{E}_{ref}, dx, dy)$
  $\mathbf{D}_p \leftarrow$ TRANSLATE3D$(\mathbf{D}_p, dx, dy, 0)$
  $dz \leftarrow$ ALIGNMEANS$(\mathbf{D}_p, \mathbf{D}_{ref})$
  $\mathbf{D}_p \leftarrow$ TRANSLATE3D$(\mathbf{D}_p, 0, 0, dz)$
// Fusion
return K-MEDIANSFUSION$(\{\mathbf{D}_p\}_{p=1...P})$

The projected DSM may have small holes due to the sampling and larger ones due to stereo mismatches. Two DSMs are produced from this projection (shown in Figure 8). In the first one the small holes are interpolated by closure with a 3×3 structuring element, while larger ones are left as no-data. This DSM will be the input of the fusion step. A second DSM is generated from the previous one, filling-in the larger holes by using the minimum value (actually the 5^{th} percentile) on the boundary of each hole. This interpolation amounts to assume that occluded parts are at ground level. This map is used for the planar alignment step.

Correlation based point cloud alignment. Because of the pointing errors in the RPC models, 3D point clouds obtained from different image pairs are usually not aligned. Bundle adjustment methods [29] simultaneously adjust the parameters of all the cameras by using image correspondences (e.g. SIFT matches [20]).

In [13] it is shown that for satellite images many of the model parameters have redundant effects, and that the affine camera model is a good approximation of the camera. This implies that displacements in the image plane are sufficient to correct the bias error using a bundle block adjustment algorithm, given enough tie-points. Still, bundle block adjustment relies on keypoint matching, which is sensitive to noise and radiometric changes, such as the ones observed in multi-date datasets, so in general large areas need to be processed in order to find enough tie-points.

In this paper we adopt a simple but effective alignment

strategy that is derived from the affine camera model and is well adapted to the case of DSMs [23]. Instead of relying on image-to-image matches it consists in matching the point clouds. This is motivated by two observations:

- Matching surface models is more stable over time than using tie-points across multi-date images (as long as the 3D geometry does not change too much);
- The error induced on the 3D point clouds by the satellite pointing error is mainly a translation [13, 23, 28].

Since the pointing error induces a 3D translation of the triangulated point clouds (see Section 2.4), we propose to align the projected DSMs by maximizing the *Normalized Cross Correlation* (NCC) between them, which is invariant to affine contrast changes. We define the NCC as

$$\mathrm{NCC}(\mathbf{u},\mathbf{v}) := \frac{1}{|\widehat{\Omega}|}\sum_{t\in\widehat{\Omega}} \frac{(\mathbf{u}_t - \mu_{\mathbf{u}}(\widehat{\Omega}))(\mathbf{v}_t - \mu_{\mathbf{v}}(\widehat{\Omega}))}{\sigma_{\mathbf{u}}(\widehat{\Omega})\,\sigma_{\mathbf{v}}(\widehat{\Omega})}, \quad (2)$$

where $\widehat{\Omega} := \Omega_{\mathbf{u}} \cap \Omega_{\mathbf{v}}$ is the intersection of the sets of known pixels in both DSMs, which allows to deal with incomplete DSMs. The sample mean and standard deviation of $\mathbf{u}$ on $\widehat{\Omega}$ are denoted respectively $\mu_{\mathbf{u}}(\widehat{\Omega})$ and $\sigma_{\mathbf{u}}(\widehat{\Omega})$.

The optimal translation aligning two DSMs $\mathbf{u}$ and $\mathbf{v}$ is determined by the maximum of the correlation

$$\mathrm{FNCC}(\mathbf{u},\mathbf{v},dx,dy) := \mathrm{NCC}(\mathbf{u},\,\mathrm{shift}(\mathbf{v},(dx,dy))), \quad (3)$$

which is maximized with a coarse-to-fine strategy.

When not aligned, no-data regions in $\mathbf{u}$ and $\mathbf{v}$ reduce the domain $\widehat{\Omega}$ where the NCC is defined. This can bias the NCC-based alignment as entire features can fall outside $\widehat{\Omega}$. To avoid this behavior the planar translation is computed using interpolated DSMs where the ground elevation is prolonged from the boundaries of missing regions ($\mathbf{E}_p$ in Algorithm 1). The altitude translation is then computed by matching the means $\mu_{\mathbf{u}}(\widehat{\Omega})$ and $\mu_{\mathbf{v}}(\widehat{\Omega})$ of the non-interpolated maps.

DSM Fusion. A popular strategy for fusing registered DSMs is the pointwise median [23]. However, the median assumes a single mode, which in a multi-date setting can yield an incoherent result due to changes in vegetation (multi-modal elevations are shown in Figure 9). To account for this multi-modality of heights we propose a method that selects the mode corresponding to the ground altitude, which is the lowest one.

We estimate the height modes at each point by applying the k-medians clustering with increasing number of clusters (1 to 8) until the clusters have a span inferior to a predefined precision. If one or two clusters are detected the lowest one is kept, otherwise the point is marked as no-data. Figure 10 compares the results obtained by the median and the proposed clustering-based strategy (denoted as k-medians).

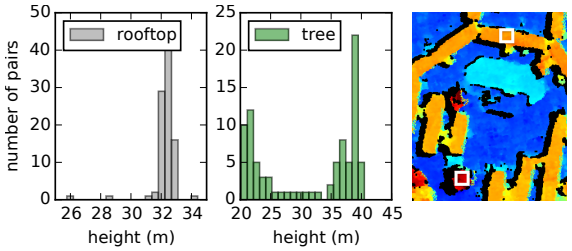


Figure 9. Taking the median for fusing multi-date DSMs is not consistent with the differences due to seasonal vegetation changes, which can be seen as bimodal.

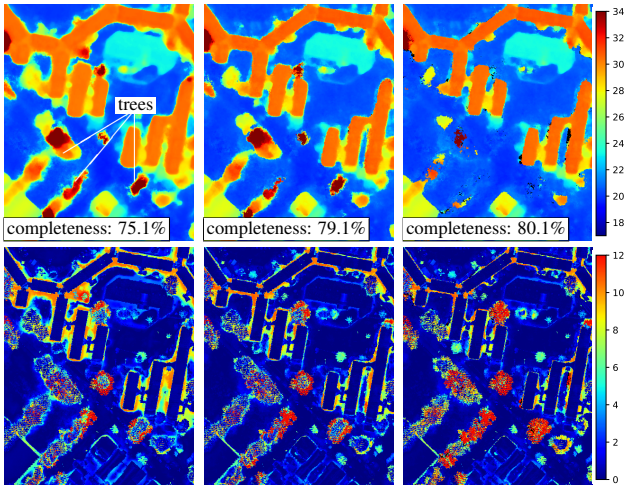


Figure 10. Median aggregation of the 700 and 50 best pairs, according to the heuristic order (left, center), and k-medians with 50 pairs (right). The second row shows the absolute difference with the ground truth. Note that the k-medians result has less foreground fattening and that many trees have disappeared.

2.4. Consequences of the Affine Camera Model

The affine camera model is a very good approximation of the pushbroom instrument [11, 13, 5, 8, 23, 28] when working on small regions of interest (e.g. about 2 km $\times$ 2 km). This approximation allows to model the pointing error by translations in the image domain [13, 7]. In this paper we take the affine camera hypothesis to another level; this implies that the triangulation function

$$\mathrm{coloc3D} : \mathbf{R}^2 \times \mathbf{R}^2 \to \mathbf{R}^3 \quad (4)$$

is linear. Thus a translation error in the image domain induces a translation in 3D. This means that we can compute the triangulation without any bundle adjustment, since any correction will result in a global 3D translation that can be easily estimated afterwards (by correlating the DSMs).

Let us formalize these observations. For an affine camera, the projection and localization functions P and L are affine maps represented as 2×3 matrices.

Definition 1 (Affine colocalization)**.** *Given a match* (p, p') *between two cameras defined by* $L = (A|b)$ *and* $L' = (A'|b')$*, the solution of the linear system* $Ap + bh = A'p' + b'h$ *for* h *is* $h = \mathrm{coloc}(p, p')$*, where*

$$\mathrm{coloc}(p, p') = \frac{(b - b') \cdot (A'p' - Ap)}{\|b - b'\|^2} \tag{5}$$

and the triangulation gives the 3D point $\mathrm{coloc3D}(p, p') = L(p, \mathrm{coloc}(p, p'))$.

Note that the affine colocalization algorithm can be applied even if the point p' does not fall on the epipolar line of p. Also note that coloc3D is a linear map (4). Now, we assume [13] that the pointing error corresponds to a translation on each image.

Lemma 2 (The bad stereo lemma)**.** *If we apply the affine colocalization algorithm to a set of matches* (p, p') *where each image domain has suffered unknown translations, we obtain a set of 3D points that differs from the correct one by a global 3D translation.*

Proof. Let us suppose that the pointing error corresponds to translations t and t' on each image. Since the function coloc3D is linear, we have

$$\mathrm{coloc3D}(p + t, p' + t') = \mathrm{coloc3D}(p, p') + \alpha_{t,t'} \tag{6}$$

where $\alpha_{t,t'} \in \mathbf{R}^3$ is a translation. □

3. Results and Discussion

We first analyze the proposed method on the training site provided with the IARPA challenge dataset [2]. Then, we validate it on three evaluation sites from the same dataset (shown in Figure 14 and rendered in Figure 2).

Training Site. Figure 11 shows the completeness and accuracy of the fused DSM as a function of the number of pairs. We compared the median [23] with the proposed k-medians fusion (Section 2.3), and evaluated three pair ordering criteria: the heuristic proposed in Section 2.1, a random order, and the *oracle order* obtained by sorting the pairs by decreasing completeness (used as reference).

The plots confirm that the proposed pair selection heuristic reaches a performance similar to the oracle, way beyond the random order. This is confirmed on the validation sites (Table 1). However, the completeness drops slightly when fusing more than 100 pairs. This motivates our choice of fusing only the first 50 pairs, instead of fusing them all (2162 in this case) as in [23]. Automatic determination of the optimal number of pairs to fuse is left for future work.

We note that the accuracy degrades with the number of fused pairs. This is justified since the pairs are ordered by decreasing completeness, so merging more pairs reduces the overall accuracy.

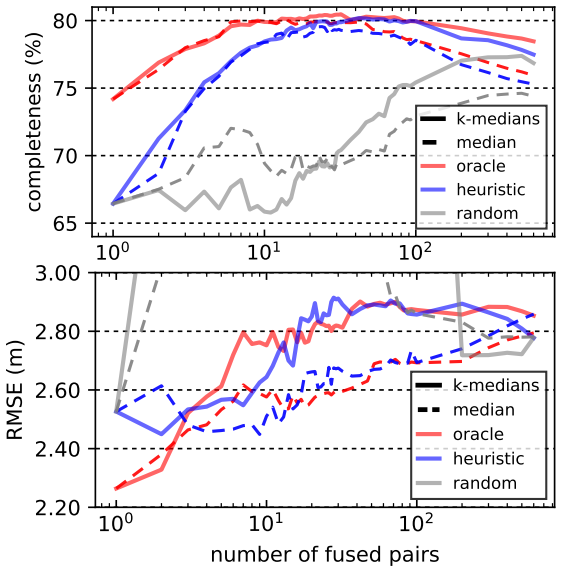


Figure 11. Completeness and accuracy (RMSE) as a function of the number of fused DSMs, where all the 47 images are used to form the pairs. Plots show the results obtained with the median and k-medians fusion for three pair ordering criteria: the proposed heuristic, a random order, and the oracle order.

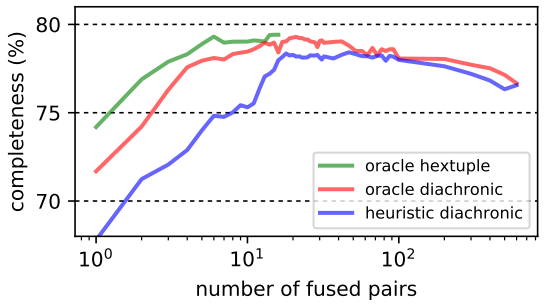
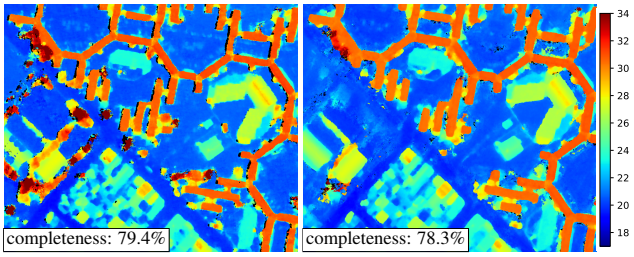


Figure 12. Completeness (training site) as a function of the number of fused DSMs, using k-medians fusion. Oracle and heuristic refer to the pair ordering criteria. The blue curve is obtained by fusing only pairs from the hextuple. The others use only pairs formed with the 38 diachronic images.

We observe that, in terms of completeness, for up until 20 pairs the k-medians fusion is similar to median [23], and only for large numbers of pairs k-medians improves over median. However, the error maps in Figure 10 show that k-medians has less foreground fattening errors, and more errors due to changes in vegetation, which degrade the accuracy. The effect of seasonal vegetation changes is the subject of future works. See more results at the project page.

We now turn to one of the main questions addressed in this paper: is it possible to obtain a quality DSM from a *diachronic* image set containing no same-date pairs? Could the result be comparable to one obtained from a same-date pair? To check this hypothesis on a meaningful example, we considered the best same-date image set, namely the hextu-



same-date hextuple (15 pairs) 38 multi-date images (50 pairs)

Figure 13. Reconstruction using only the hextuple of same-date images (left), and 38 diachronic images (right). The first one is obtained by fusing 15 pairs, the second is a fusion of 50 pairs selected with the heuristic criterion. Note the differences in the trees (dark red).

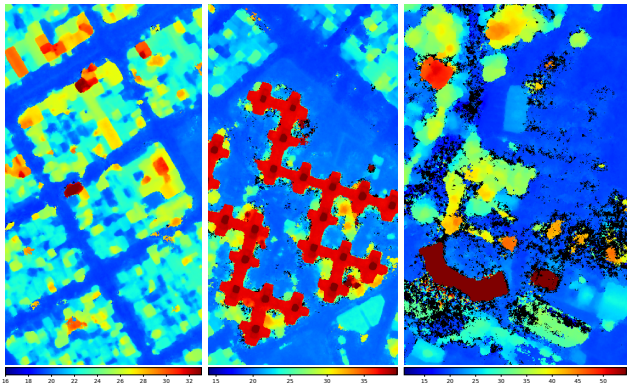


Figure 14. Results on three evaluation sites, fusing with k-medians the 50 best pairs selected by the heuristic criterion. High resolution results and difference maps with ground truth are available at the project webpage.

ple provided in the IARPA dataset. We fused its pairs using the oracle order. This represents one of the best results obtainable with a rich same-date image set, here composed of six images. Then, we formed pairs using only the 38 diachronic images, and fused them according to the oracle and heuristic order. The curves in Figure 12 show that, using only diachronic images, it is possible to attain the quality of a same-date set. We note that the completeness obtained with the diachronic set rapidly surpasses a single same-date pair, and that the difference is about 1% when compared to the full hextuple. Figure 13 shows both resulting DSMs; note that some trees have disappeared in the diachronic result due to the k-medians fusion.

Evaluation Sites. We applied the proposed method on the three evaluation sites provided with the IARPA dataset [2]. The sites, shown in Figure 14, have different characteristics: low and medium-rise buildings with few trees in sites 1 and 2, and high-rise buildings with many trees in site 3. In addition, sites 2 and 3 are not seen in 7 of the 47 images, including the same-date hextuple. Thus the method has to cope with fewer images. Each site depicts an area of about 400×400 meters at 30 cm.

	heuristic order		oracle order	
site	**med**	**k-med**	**med**	**k-med**
training	79.0 / **2.67**	**80.1** / 2.89	79.3 / **2.69**	**80.2** / 2.89
site 1	73.6 / **1.80**	**74.0** / 1.88	74.4 / **1.79**	**74.7** / 1.88
site 2	71.8 / 3.97	**73.1** / **3.87**	71.6 / 3.85	**73.1** / **3.79**
site 3	57.2 / **6.73**	**58.6** / 7.52	57.9 / **6.36**	**59.6** / 6.98

Table 1. Completeness (%) / Accuracy (m) of fused DSMs using 50 pairs. We compare the heuristic and oracle (from training) pair selection, and the median (**med**) and k-medians (**k-med**) fusion.

The results of fusing 50 pairs with the median and k-medians strategies are presented in Table 1, while Figure 14 illustrates the k-medians result computed using the heuristic pair selection. The completeness drop of site 3 is due to a higher vegetation density on this site. As for the training site, the k-medians shows a small improvement compared to the median. But the results have less foreground fattening.

Last, since the proposed method only computes a fraction of all the image pairs, its computational cost is one order of magnitude lower than [23], which matches all the pairs. Each site is computed in less than 1 hour on a 16-core computer.

4. Conclusion

We propose an algorithm to compute a 3D reconstruction from a collection of satellite images of the same site. The method is able to add information from new images incrementally and it does not rely on a global bundle adjustment. It relies instead on the local affine camera approximation [13], which allows to compute 3D models independently from the original pairs of images, then aligns the models by 3D translations. Experiments show that a 3D model computed by our algorithm from a multi-date collection can be as accurate as a 3D model computed from a pair of same-date images. We propose a heuristic to select the best image pairs from a large collection, and we observe that the optimal result is obtained by keeping only few well-chosen pairs from the large set of all possible pairs. Finally, since DSMs often exhibit a yearly oscillation (due to deciduous trees), we propose a fusion criterion that gives a "winter" version of the DSM. Our experiments rely on the recently published IARPA dataset [2], which proved an invaluable tool to assess the validity of the proposed techniques. Future work will focus on the evaluation on more datasets and the comparison between classic bundle adjustment and the proposed DSM fusion strategies.

Acknowledgements ONR grant N00014-14-1-0023, CNES, MISS Project and ANR-12-ASTR-0035 DGA project. The authors would like to thank Jean-Michel Morel for his help and fruitful discussions.

References

[1] K. J. Becker, B. A. Archinal, T. H. Hare, R. L. Kirk, E. Howington-Kraus, M. S. Robinson, and M. R. Rosiek. Criteria for Automated Identification of Stereo Image Pairs. *46th Lunar and Planetary Science Conference, held March 16-20, 2015 in The Woodlands, Texas. LPI Contribution No. 1832, p.2703*, 46:2703, 2015. 3

[2] M. Bosch Ruiz, Z. Kurtz, H. Shea, and M. Brown. A Multiple View Stereo Benchmark for Satellite Imagery. In *Proceedings of the IEEE Applied Imagery Pattern Recognition (AIPR) Workshop*, 2016. 1, 2, 4, 5, 7, 8

[3] P. d'Angelo. Improving Semi-Global Matching: Cost Aggregation and Confidence Measure. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B1:299–304, 6 2016. 5

[4] P. d'Angelo and G. Kuschk. Dense multi-view stereo from satellite imagery. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 6944–6947. IEEE, 7 2012. 3, 4

[5] P. d'Angelo and P. Reinartz. DSM based orientation of large stereo satellite image blocks. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXIX-B1:209–214, 7 2012. 4, 6

[6] C. de Franchis, E. Meinhardt-Llopis, J. Michel, J.-M. Morel, and G. Facciolo. An automatic and modular stereo pipeline for pushbroom images. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume II, pages 49–56, Zurich, 8 2014. 3, 4

[7] C. de Franchis, E. Meinhardt-Llopis, J. Michel, J.-M. Morel, and G. Facciolo. Automatic sensor orientation refinement of Pléiades stereo images. In *Geoscience and Remote Sensing Symposium (IGARSS), 2014 IEEE International*, pages 1639–1642, Québec, 2014. 6

[8] C. de Franchis, E. Meinhardt-Llopis, J. Michel, J.-M. Morel, and G. Facciolo. On stereo-rectification of pushbroom images. In *Proceedings of the International Conference on Image Processing (ICIP)*, 2014. 1, 6

[9] G. Dial and J. Grodecki. RPC replacement camera models. *Proc. ASPRS Annual Conference, Baltimore*, pages 1–5, 2005. 2

[10] G. Facciolo, C. de Franchis, and E. Meinhardt-Llopis. MGM: A Significantly More Global Matching for Stereovision. In *Procedings of the British Machine Vision Conference 2015*, pages 1–90. British Machine Vision Association, 2015. 5

[11] W. Förstner. Quality Assessment of Object Location and Point Transfer Using Digital Image Correlation Techniques. In *ISPRS Congress XXV, Rio de Janeiro, Brasil*, pages 1–23, 1984. 1, 6

[12] Y. Furukawa and C. Hernández. Multi-View Stereo: A Tutorial. *Foundations and Trends in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 1

[13] J. Grodecki and G. Dial. Block adjustment of high-resolution satellite images described by Rational Polynomials. *Photogrammetric Engineering and Remote Sensing*, 69(1):59–68, 2003. 1, 2, 3, 5, 6, 7, 8

[14] C. Guerin, R. Binet, and M. Pierrot-Deseilligny. Automatic Detection of Elevation Changes by Differential DSM Analysis: Application to Urban Areas. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(10):4020–4037, 10 2014. 3

[15] I. Haller, C. Pantilie, F. Oniga, and S. Nedevschi. Real-time semi-global dense stereo solution with improved sub-pixel accuracy. *IEEE Intelligent Vehicles Symposium, Proceedings*, pages 369–376, 2010. 5

[16] H. Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–41, feb 2008. 5

[17] H. Hirschmüller and S. Gehrig. Stereo matching in the presence of sub-pixel calibration errors. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, pages 437–444, 2009. 5

[18] J. J. Koenderink and a. J. van Doorn. Affine structure from motion. *Journal of the Optical Society of America. A, Optics and image science*, 8(2):377–385, 1991. 4

[19] G. Kuschk. Large scale urban reconstruction from remote sensing imagery. In *3D-ARCH 2013 - 3D Virtual Reconstruction and Visualization of Complex Architectures*, volume XL-5/W1, pages 139–146, Trento, feb 2013. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 1, 3, 4

[20] D. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 91–110, 1999. 5

[21] Z. Moratto, O. Alexandrov, S. McMichael, and R. Beyer. *The Ames Stereo Pipeline: NASA's Open Source Automated Stereogrammetry Software*. NASA, 2014. 1, 4

[22] Z. M. Moratto, M. J. Broxton, R. A. Beyer, M. Lundy, and K. Husmann. Ames Stereo Pipeline, NASA's Open Source Automated Stereogrammetry Software. *41st Lunar and Planetary Science Conference, held March 1-5, 2010 in The Woodlands, Texas. LPI Contribution No. 1533, p.2364*, 41:2364, 2010. 1

[23] O. C. Ozcanli, Y. Dong, J. L. Mundy, H. Webb, R. Hammoud, and V. Tom. A comparison of stereo and multiview 3-D reconstruction using cross-sensor satellite imagery. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 17–25. IEEE, 6 2015. 1, 2, 3, 6, 7, 8

[24] O. C. Ozcanli, Y. Dong, J. L. Mundy, H. Webb, R. Hammoud, and V. Tom. Automatic Geolocation Correction of Satellite Imagery. *International Journal of Computer Vision*, 116(3):263–277, 2 2016. 1

[25] M. Pierrot Deseilligny. MicMac, Apero, Pastis and Other Beverages in a Nutshell. *MicMac Documentation*, 2015. 4

[26] T. B. Pollard, I. Eden, J. L. Mundy, and D. B. Cooper. A Volumetric Approach to Change Detection in Satellite Images. *Photogrammetric Engineering & Remote Sensing*, 76(7):817–831, 7 2010. 1, 2

[27] H. Schmid. An analytical treatment of the problem of triangulation by stereophotogrammetry. *Photogrammetria*, 13:67–77, 1 1956. 1

[28] D. E. Shean, O. Alexandrov, Z. M. Moratto, B. E. Smith, I. R. Joughin, C. Porter, and P. Morin. An automated, open-source pipeline for mass production of digital elevation models (DEMs) from very-high-resolution commercial stereo satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 116:101–117, 6 2016. 1, 3, 6

[29] T. Toutin. Review article: Geometric processing of remote sensing images: models, algorithms and methods. *International Journal of Remote Sensing*, 25(10):1893–1924, 5 2004. 5

[30] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle Adjustment — A Modern Synthesis. In *Vision Algorithms '99*, volume 34099, pages 298–372. 2000. 1

[31] J. Wohlfeil, H. Hirschmüller, B. Piltz, A. Börner, and M. Suppa. Fully automated generation of accurate digital surface models with sub-meter resolution from satellite imagery. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXIX-B3(September):75–80, jul 2012. 4

[32] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In *Computer Vision — ECCV '94*, number May, pages 151–158. Springer-Verlag, Berlin/Heidelberg, 1994. 5