To Bundle Adjust or Not: A Comparison of Relative Geolocation Correction Strategies for Satellite Multi-View Stereo

Roger Marí Carlo de Franchis Enric Meinhardt-Llopis Gabriele Facciolo
CMLA, ENS Paris-Saclay, France
https://rogermm14.github.io/bundle-adjust-or-not

Abstract

The generation of up-to-date accurate 3D models from multi-view satellite images has recently become a hot research topic. A well-known challenge of this problem is to put all cameras into a common frame of reference, since depending on the satellite geopositioning equipment the camera parameters may contain errors of up to tens of meters on the ground. In this context, bundle adjustment based techniques, relying on the identification of a set of tie-points and the correction of the camera models to make them coincident, have become a generally accepted practice. However, new approaches capable of producing state-of-the-art results without the use of prior bundle adjustment have also been proposed. This work aims to compare both strategies and assess the practical impact of using bundle adjustment for 3D reconstruction from multi-view satellite images.

1. Introduction

In the past few years, the advances in satellite technology have resulted in a remarkable increase of high resolution imagery of the Earth surface, with many areas being captured on a daily basis or multiple times per year. Current satellite imagery allows the use of photogrammetry to build accurate 3D digital surface models (DSMs) in a periodic manner, providing extremely valuable up-to-date information about the evolution of the terrain and the human activity on it. The applications of this field are multiple and ambitious, including navigation, urban planning, surveillance or natural disaster prevention and response.

Compared to multi-view satellite images, airbone lidar acquisitions and aerial images are well-known alternatives that can be used to create 3D models of higher precision. However, these technologies are limited by their narrow swath and cost, making it difficult to update the models in a short period of time. In this context, the break-through of low-cost and high resolution satellite imagery has enabled persistent coverage of large areas, inaccessible or not.

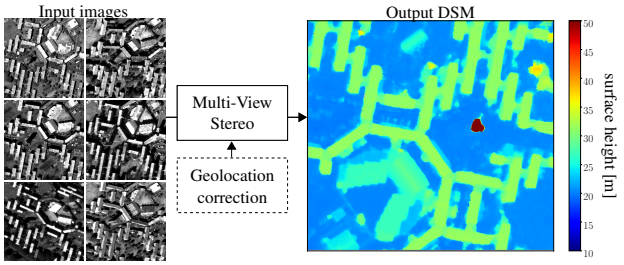

Figure 1: Geolocation correction methods are a key step to remove the effects of pointing errors and thus obtaining precise output models in 3D reconstruction from multi-view satellite images.

Satellite images are typically provided along with a Rational Polynomial Coefficients (RPC) camera model [8], and other metadata such as the acquisition timestamp or the pixel size. The RPC camera model is composed by two rational polynomial functions that approximate the mapping from 3D space points to 2D image pixels (i.e. the *projection* function) and its inverse (i.e. the *localization* function). RPCs allow to model complex camera systems independently of the specifics on the system (e.g. pushbroom or projective). Given a stereo correspondence found across two or more satellite images, the associated RPCs can be used to triangulate and retrieve the 3D point that projects on the correspondence.

Although RPCs are expected to be precise enough, the complex system they encode is subject to measurement errors in the satellite geopositioning equipment, mainly due to the attitude angles. Such inaccuracies, also referred to as *pointing errors* [22], can be of the order of tens of pixels in the image domain. This implies that different satellite views are typically not consistent in a common frame of reference (i.e. each 3D point projects to a slightly different location in the images). Hence it is imperative to use some relative geolocation correction strategy to prevent these pointing errors from affecting the triangulation of 3D points and the accuracy of the output DSMs.

Techniques for 3D reconstruction from multiple views can be grouped into two categories:

– *True multi-view* methods tackle the multi-view triangulation problem for all images simultaneously.

– *Multi-view stereo* (MVS) methods process image pairs independently and then fuse the resulting 3D models.

MVS methods are the most popular choice for satellite imagery, as it has been shown that they can outperform sophisticated true multi-view techniques [21]. The efficient and stable performance of the semi-global matching (SGM) algorithm [16] typically used in stereo pipelines (e.g. [7, 11, 23, 9, 4] among others) is usually pointed out as the driving force behind their success.

The objective of this work is to investigate the performance of 3 MVS pipelines for 3D reconstruction from multi-date satellite images. Two of the pipelines reviewed in this paper use a bundle adjustment procedure, each optimizing a different set of parameters to correct the RPCs: 2D translations in the image domain [22] or the 3D rotation of the satellite cameras [25]. Oppositely, the third approach omits any kind of prior bundle adjustment and aligns independent models based on geometry correlation [11]. To perform a fair comparison, the different blocks of the MVS pipelines are all common apart from the relative geolocation correction method that is applied in each case.

Our evaluation will focus on the reconstruction of small areas of interest. Indeed, the increasing availability of satellite imagery has enabled the exploitation of incidental imagery for 3D reconstruction [18]. This is useful for monitoring applications in which a concrete area needs to be reconstructed from the available imagery. Moreover, modern satellite constellations, such as SkySat from Planet, favor smaller footprints and revisit capacity to large swaths.

We employ the *IARPA Multi-View Stereo 3D Mapping Challenge* dataset [2], which includes 47 DigitalGlobe WorldView-3 images, with 30 cm nadir resolution, collected between 2014 and 2016 over Buenos Aires; and the dataset comprised in the *2019 IEEE GRSS Data Fusion Contest* [17, 1] with 26 DigitalGlobe WorldView-3 images collected between 2014 and 2016 over Jacksonville. The completeness (percentage of points where the absolute difference is less than 1 meter) of the output models with respect to the lidar ground truth is used as the main evaluation metric to assess the performance of the different methods.

1.1. Bundle adjustment

Given a cloud of K 3D points $\{\boldsymbol{X}_k\}_{k=1,\dots,K}$, a set of feature tracks representing their projections across M images, and the projection maps $\{P_m\}_{m=1,\dots,M} : \mathbf{R}^3 \to \mathbf{R}^2$ of the cameras (usually associated to a projection matrix or a RPC model), bundle adjustment is the process that seeks to minimize the reprojection error of the setting by optimizing $\{\boldsymbol{X}_k\}$ and $\{P_m\}$. The reprojection error is defined as the sum of the squared Euclidean distances between the projections of the point cloud $\hat{\boldsymbol{x}}_{mk} = P_m(\boldsymbol{X}_k)$ and the real measurements of $\boldsymbol{x}_{mk}$ (i.e. the detected features):

$$E(\{P_m\}, \{\boldsymbol{X}_k\}) = \sum_{k=1}^{K} \sum_{m=1}^{M} \|\boldsymbol{x}_{mk} - P_m(\boldsymbol{X}_k)\|^2. \quad (1)$$

1.2. Related work

The Computer Vision community has proposed different methods to correct the pointing error of RPC camera models. Bundle adjustment based solutions are a generally accepted practice that consist in detecting inter-image tie-points and applying a compensating function to the original RPCs so that the back-projections of the tie-points are coincident in the 3D world [14, 5, 22, 13, 18].

In the case of satellite images, since satellites are far from the Earth's surface, the main component of the reprojection error comes from the inaccurate knowledge of the satellite orientation. This means that the energy in Equation 1 can be minimized by composing each P_m with a global 3D transformation R_m. This amounts to finding the $\{R_m\}$ that minimize $\sum_k \|P_m(R_m(\boldsymbol{X}_k)) - \boldsymbol{x}_{mk}\|^2$.

In [14] it was shown that the net effect of pointing error reduces to a 2D translation (also termed as *bias* or *correction offset*) in images accounting for less than 50 km in each dimension. Based on this observation, *bundle block adjustment* procedures have in common the optimization of 2D correction offsets. Following the notation from Equation 1, this amounts to finding the M 2D transformations $\{T_m\}$ that minimize $\sum_k \|T_m(P_m(\boldsymbol{X}_k)) - \boldsymbol{x}_{mk}\|^2$. Note that this problem is easier than the previous one as it reduces to a linear system. Also note that bundle adjustment and bundle block adjustment correspond to inserting a correction before the projection or after it.

For smaller areas (e.g. 2 km $\times$ 2 km), RPCs can be modeled as affine cameras using the first order Taylor approximation [12, 24, 11, 6]. This allows to correct the effect of bias on the triangulated points with an affine 3D transformation. After this idea, [11] proposed a new approach capable of producing high-quality reconstructions without needing a bundle adjustment. Independent DSMs are registered based on correlation. Hence geometry is used instead of image information to solve the problem of relative geolocation correction. The main motivation behind this work is that finding a sufficient amount of tie-points can be an issue with multi-date images, especially when restricted to small areas of interest. Significant differences due to noise or radiometric changes may cause image matches not to be accurate enough, set aside the impact of human activity, weather phenomena or seasonal changes.

Our work tries to give a deeper vision on the subject, highlighting the pros and cons of image (i.e. bundle adjustment) and geometry based solutions and revealing aspects to take into account when applying them.

2. Methodology

This section presents the most relevant concepts behind the 3 MVS pipelines for 3D reconstruction from multi-date satellite images reviewed in this study, focusing on the relative geolocation correction methods proposed by each.

2.1. Correlation based DSM alignment

Originally proposed in [11], this pipeline aggregates 3D point clouds independently computed from N different stereo pairs, without any prior RPC correction. The method is summarized in Algorithm 1. Our implementation is based on the publicly available satellite stereo pipeline S2P [7, 3].

The input images are cut into tiles covering small areas, where RPCs can be locally approximated as an affine camera model (see Section 1.2). The reconstruction of each tile starts by rectifying the image crops (I, I') of each stereo pair, to make epipolar lines horizontal. A variant of the SGM algorithm [10] with a cost based on the census transform is then used to compute a disparity map robust to lighting changes. Only consistent disparities passing the left-right check are kept. The correspondences given by the disparity map are re-expressed in the original images domain and triangulated using the affine projection matrices P, P' of the pair to compute a dense 3D point cloud.

After running the previous process for the N input stereo pairs, the objective is to merge the output point clouds to obtain a high-quality reconstruction. To this end, each point cloud is projected on a geographic grid, thus producing different DSMs as in [2, 11]. Morphological filters are then used to refine the DSMs, which may contain small holes due to the sampling step and larger ones due to mismatches. Closing with a 3×3 structuring element is applied to fill small holes, followed by interpolation using a low value (the 5th percentile) on the boundaries to reduce larger holes. This interpolation strategy assumes that occluded parts are at ground level.

At this point, the post-processed DSMs are not aligned due to the pointing error in the satellite RPCs, which prevents any kind of fusion. This is where the differential part of the pipeline intervenes: the DSMs are aligned via a 3D translation that maximizes the Normalized Cross Correlation (NCC) between them, defined 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 points in two DSMs $\mathbf{u}$ and $\mathbf{v}$. The mean and standard deviation of $\mathbf{u}$ on $\hat{\Omega}$ are denoted respectively $\mu_{\mathbf{u}}(\hat{\Omega})$ and $\sigma_{\mathbf{u}}(\hat{\Omega})$.

According to [11], this is motivated by two observations:

- The misalignment induced by the satellite pointing error is mainly a translation [14, 12, 24, 6].
- As long as the 3D geometry of the area does not change too much, matching surface models is more stable over time than using tie-points across multi-date images.

Algorithm 1: *MVS with NCC based DSM alignment*

Input : M views of a small area of interest (AOI) cropped from multi-date satellite images + associated RPCs
Output: High-quality DSM of the input AOI

- Select N stereo pairs: $\{ (I_n, I'_n) \}_{n=1,\dots N}$
for each stereo pair (I_n, I'_n) **do**
  - Affine approx. of raw RPCs $\rightarrow P_n, P'_n$
  - Epipolar rectification
  - Dense stereo matching
  - Triangulate using P_n, P'_n to get 3D point cloud
  - Project point cloud to DSM_n
  - Post-process DSM_n
end
- DSM alignment via 3D translations maximizing the NCC
- Fuse all DSMs via point-wise median

The maximum correlation translations are employed to register all DSMs to the frame of reference of the first input stereo pair, which is expected to be the best according to the selection criterion used (see Section 3.1). After the alignment, the point-wise median is used to perform the DSMs fusion, as in [13]. Remark that the fusion is done using the DSMs previous to interpolation. Therefore, in practice, the interpolated DSMs are only used to compute the alignment transformations. This is done to avoid possible biases due to large areas of unknown values, as detailed in [11]).

2.2. Bundle block adjustment

This MVS pipeline (Algorithm 2), based on [22], was selected to test a bundle block adjustment procedure. The RPCs are corrected previous to the triangulation of stereo correspondences, implying that all DSMs are aligned before the fusion step. The approach aims to find the 2D translations (or correction offsets) in the image domain that compensate the pointing error of each view (see Figure 2).

The first step of the pipeline consists in the detection of feature tracks across the set of input images. The feature tracks employed in our experiments result from pairwise matches of SIFT keypoints [19]. We apply a distance ratio test as in [19] with a rather strict threshold of 0.6, and also perform geometric filtering using the Fundamental matrix to minimize the presence of outliers. The union-find algorithm from [20] is employed to extend pairwise matches to unordered tracks of arbitrary length in an efficient way.

The feature tracks are used to initialize a sparse 3D point cloud and a correction offset for each image, providing this way the necessary inputs for the bundle block adjustment. Algorithm 3, introduced in [22], details how to initialize the correction offsets and the 3D points from the feature tracks.

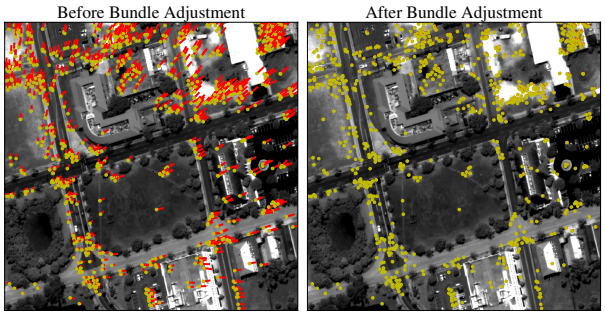

Figure 2: Effect of pointing error before and after bundle adjustment. Green dots represent the detected features, and red vectors the distance to the reprojected locations (i.e. reprojection error). After bundle adjustment, the reprojection error reaches subpixel magnitude, implying that the RPCs have been corrected.

Note that initializing the values for the 3D points is not a straight-forward task: the direct triangulation of each stereo correspondence part of a track would produce a different 3D point because of the still to be corrected pointing error.

Using the RPC localization functions, it is possible to back-project a ray from each track observation to a set of horizontal planes with heights $\{Z_{\min}, ..., Z_{\max}\}$, given by a series of ΔZ increments, covering a sufficient range to contain the whole scene. The height Z where the multiple back-projections are less scattered (i.e. minimum σ_Z, as defined in Equation 3) defines the initial depth of the 3D point associated to the track. The (X, Y) coordinates are given by the mean (μ_X, μ_Y) of all back-projections at height Z.

The scatter value σ_Z at height Z for a given feature track is defined as

$$\sigma_Z = \sqrt{\sum_i (X_i - \mu_X)^2 + \sum_i (Y_i - \mu_Y)^2}, \qquad (3)$$

where X_i and Y_i are the coordinates of the back-projection of the i-th observation of the feature track at height Z.

To initialize the correction offsets of all input images, all possible offsets per image are computed using each feature track. Observe that error-free tracks should ideally generate the same offsets for all images if the model holds. Adaptive RANSAC can be applied then to pick a single correction offset per image with the largest consensus [15]. RANSAC threshold to declare inliers was set to 3 pixels. Further refinement of the tracks and the initial location of the 3D points is done in [22] by preserving, for each image, only those observations that contributed to an inlier offset. For simplicity, we kept all raw feature tracks and 3D points as output by Algorithm 3 to feed the bundle block adjustment.

As mentioned in Section 1, apart from the relative geolocation correction method, the rest of the blocks of the approach do not change in comparison to the other MVS pipelines reviewed in this study (i.e. steps of lines 10–19 in Algorithm 2 are the same in Algorithm 1).

Algorithm 2: *MVS with bundle block adjustment*

Input : M views of a small area of interest (AOI) cropped from multi-date satellite images + associated RPCs
Output: High-quality DSM of the input AOI

1 - Feature track detection across the M input images
2 - Run Algorithm 3 to compute:
3 (1) All possible correction offsets for all images
4 (2) An initial value for the 3D point of each track,
5 i.e. point cloud X
6 **for** each image I_m **in** $\{I_1, ..., I_M\}$ **do**
7 - RANSAC to select an offset $\boldsymbol{\rho}_m$ with a large support
8 **end**
9 - Bundle block adjustment to refine the M correction offsets and $X \rightarrow \{\boldsymbol{\rho}_{\text{BA}_m}\}_{m=1,...M}$, X_{BA}
10 - Select N stereo pairs: $\{(I_n, I'_n)\}_{n=1,...N}$
11 **for** each stereo pair (I_n, I'_n) **do**
12 - Affine approx. of corrected RPCs $\rightarrow P_{\text{BA}_n}, P'_{\text{BA}_n}$
13 - Epipolar rectification
14 - Dense stereo matching
15 - Triangulate using $P_{\text{BA}_n}, P'_{\text{BA}_n}$ to get 3D point cloud
16 - Project point cloud to DSM_n
17 - Post-process DSM_n
18 **end**
19 - Fuse all DSMs via point-wise median

Algorithm 3: *Correction offsets from feature tracks*

Input : M views of a small area of interest (AOI) cropped from multi-date satellite images + associated RPCs
Feature track k detected across the input images
Output: M correction offsets, one per image
3D point $\boldsymbol{X}_k$ corresponding to feature track k

1 **for** each altitude Z **in** $\{Z_{\min}, ..., Z_{\max}\}$ **do**
2 **for** each image I_m **in** $\{I_1, ..., I_M\}$ **do**
3 - Pick the 2D observation of track k in $I_m \rightarrow \boldsymbol{x}_{mk}$
4 - Localize $\boldsymbol{x}_{mk}$ at height Z via $\text{RPC}_m \rightarrow (X, Y, Z)$
5 - Add (X, Y, Z) to the list of 3D point candidates for track k at height $Z \rightarrow \text{LIST}_k$
6 **end**
7 - Compute σ_Z of LIST_k as stated in Equation 3
8 **end**
9 - Define the 3D point of track k as $\boldsymbol{X}_k = \min_{\sigma_Z} \text{LIST}_k$
10 **for** each image I_m **in** $\{I_1, ..., I_M\}$ **do**
11 - Project $\boldsymbol{X}_k$ via $\text{RPC}_m \rightarrow \hat{\boldsymbol{x}}_{mk}$
12 - Compute the correction offset $\rightarrow \boldsymbol{\rho}_{mk} = \hat{\boldsymbol{x}}_{mk} - \boldsymbol{x}_{mk}$
13 **end**

2.3. Bundle adjustment of camera rotations

This pipeline is presented as an alternative to traditional bundle block adjustment, without renouncing to the correction of the RPCs. Instead of optimizing a set of correction offsets, bundle adjustment is used to correct the orientation (i.e. rotation) of the satellite cameras and compensate the pointing error. The approach is outlined in Algorithm 4.

Algorithm 4: *MVS with correction of camera rotation*

Input : M views of a small area of interest (AOI) cropped from multi-date satellite images + associated RPCs
Output: High-quality DSM of the input AOI

- Feature track detection across the M input images
- Affine approx. of all raw RPCs $\rightarrow \{ P_m \}_{m=1,\dots M}$
- Initialize sparse point cloud from feature tracks $\rightarrow X$
- Bundle adjustment to refine the M rotation matrices and $X \rightarrow \{ P_{\text{BA}_m} \}_{m=1,\dots M}$, X_{BA}
- Select N stereo pairs: $\{ (I_n, I'_n) \}_{n=1,\dots N}$
for each stereo pair (I_n, I'_n) **do**
    - Epipolar rectification
    - Dense stereo matching
    - Triangulate using $P_{\text{BA}_n}, P'_{\text{BA}_n}$ to get 3D point cloud
    - Project point cloud to DSM_n
    - Post-process DSM_n
end
- Fuse all DSMs via point-wise median

As in the previous MVS pipelines, since the area to reconstruct is assumed to be small, the RPCs can be locally modeled as affine camera projection matrices. The affine camera model can be decomposed as

$$P_{3\times 4} = \begin{pmatrix} M_{2\times 3} & \boldsymbol{t}_{2\times 1} \\ 0 & 1 \end{pmatrix}, \tag{4}$$

where $M_{2\times 3} = K_{2\times 2} R_{2\times 3}$, being K the calibration matrix and R and $\boldsymbol{t}$ the camera rotation and position respectively. The method aims to refine the R matrix of each camera.

Note that a small rotation of a camera far away from a scene, as it is the case for satellite imagery, amounts in practice to a translation on the image domain [14]. This observation suggests that optimizing the rotation matrices R with bundle adjustment should be, at least, equivalent to the traditional offset correction. The rest of parameters of the affine projection matrices are fixed.

Since it is known that reducing the number of parameters to be optimized aids the bundle adjustment process, we encode all rotation matrices using the Euler angles as a 3-parameter representation. Any 3D rotation matrix R can be decomposed into 3 elemental rotations with respect to the world reference system, where the Euler angles ϕ, θ, α are the angles of rotation around the canonical axes:

$$\begin{aligned} R &= R_x(\phi) R_y(\theta) R_z(\alpha) \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & -\sin\phi \\ 0 & \sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix} \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}. \end{aligned} \tag{5}$$

In Algorithm 4, feature tracks are detected following the same methodology from Section 2.2. Differently from Algorithm 3, the 3D points associated to the tracks are initialized by triangulating all pairwise matches per track and taking the mean of the resulting 3D locations, which is faster than the technique used in Algorithm 3.

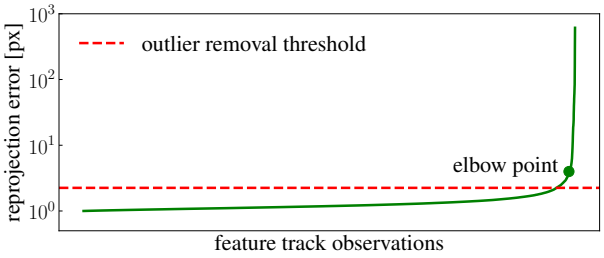

Figure 3: Feature track refinement. After an initial bundle adjustment with soft L^1 loss, sorting the reprojection errors typically results into a well-defined elbow-shaped function (due to the large errors caused by a small subset of outliers). We set an outlier removal threshold to the 95th percentile of all values below the elbow point (i.e. the point with largest distance with respect to the line defined by the minimum and maximum errors).

2.4. Feature track refinement

It should be noted that the result of the bundle adjustment based approaches is highly dependent on the quality of the input feature tracks found across the input images.

The classic bundle adjustment loss considers the squared distances between projected and observed locations (i.e. $L^2(d) = d^2$, where d denotes the Euclidean distance), since the L^2 norm is well-posed for differential calculus and the optimization converges rapidly. Nonetheless, this loss is very sensitive to the presence of outliers, which can cause the result to be biased according to erroneous tracks at the expense of good observations.

As a solution, a combination of the L^1 and L^2 losses such as the soft L^1 loss can be used. The soft L^1 loss is defined as

$$L^1_{\text{soft}}(d) = 2\left(\sqrt{1+d^2} - 1\right), \tag{6}$$

where d is the Euclidean distance between two points.

The soft L^1 loss from Equation 6 offers higher robustness to outliers. It behaves as a linear loss for large distances, which are likely to be caused by outliers; and as a quadratic loss for smaller distances around 1 pixel or below, likely to be caused by inliers.

Based on to the previous, we employ the following procedure to improve the quality of the feature tracks. We run two successive bundle adjustment steps: the first one uses the soft L^1 loss and the second one the L^2 loss. Thanks to the soft L^1 loss, after the first run we can expect the gap in terms of reprojection error between inlier and outlier observations to increase. As shown in Figure 3, a threshold can be set to discard erroneous observations according to this error. The remaining tracks (presumably made of reliable inliers) can be fed to the second run, using the L^2 loss, which yields the optimal estimator for Gaussian perturbations and quickly converges to a refined solution.

	IARPA	JAX 113	JAX 161	JAX 251
Oracle order				
Correlation based DSM alignment	70.62 / 2.67	–	–	–
Bundle block adjustment - *naif*	64.39 / 2.72	–	–	–
Bundle block adjustment	70.63 / 2.74	–	–	–
Bundle adjustment of camera rotations - *naif*	64.50 / 2.71	–	–	–
Bundle adjustment of camera rotations	70.71 / 2.74	–	–	–
Heuristic order				
Correlation based DSM alignment	68.08 / 2.69	77.72 / 2.00	82.75 / 1.70	74.87 / 2.90
Bundle block adjustment	69.73 / 2.74	77.74 / 2.04	82.53 / 1.73	76.86 / 2.91
Bundle adjustment of camera rotations	69.89 / 2.75	77.72 / 2.04	82.60 / 1.72	75.91 / 2.91
SIFT order				
Correlation based DSM alignment	48.84 / 2.62	76.73 / 2.01	82.64 / 1.64	72.46 / 2.74
Bundle block adjustment	42.15 / 2.71	76.83 / 2.06	82.48 / 1.66	72.69 / 2.76
Bundle adjustment of camera rotations	42.15 / 2.71	76.79 / 2.04	82.44 / 1.66	71.14 / 2.78

Table 1: Completeness (%) / Accuracy (m) of the reconstructed DSMs for the IARPA (Buenos Aires, one AOI) and GRSS (Jacksonville, AOIs 161, 251 and 113) datasets. The *naif* label indicates that a single bundle adjustment run with classic L^2 loss for reprojection errors was used. Otherwise, the feature track refinement strategy from Section 2.4 was used in bundle adjustment procedures.

3. Evaluation

This section presents the conducted experiments and the results of our study. Table 1 summarizes the performance metrics for an area of interest (AOI) from the IARPA dataset and three AOIs from the GRSS dataset respectively. Examples of output DSMs are shown in Figure 4. Note that the reconstruction may contain unknown values, represented as white points in Figure 4, if no stereo pair finds a reliable correspondence for certain areas.

The reconstructed DSMs and the ground truth DSMs of each site may not be in the same frame of reference. We employ a translation that maximizes the correlation between both models to register them, following the procedure from Section 2.1. After this, the performance metrics are computed from the error between the two surfaces, that is defined as the point-wise absolute difference (in meters).

Completeness represents the percentage of points whose error is less than 1 meter, with unknown values being counted as larger errors. The accuracy value, in meters, is the root mean square error (RMSE) of all known points. Both metrics are defined in [2]. Points within water bodies were not taken into account.

3.1. Selection of input pairs

Previous work already highlighted the importance of the criterion used to select input stereo pairs for MVS pipelines dealing with satellite imagery [11, 13]. Poor choices lead to pairs of views sharing less visual content and output DSMs with larger errors and incomplete areas, making the correction of the pointing error harder both for image and geometry based methods. We ran the MVS pipelines using 3 different criteria to assess the robustness of the geolocation correction methods:

- **Oracle order:** Obtained by computing the DSM of each possible stereo pair and sorting the pairs by decreasing completeness. It guarantees that the best pairs are selected, but it is expensive to compute and unrealistic since ground truth may not be available and is needed to compute the completeness of each pair.
- **Heuristic order:** Detailed in [11], this criterion is an attempt to emulate the oracle order based on the metadata of the satellite images. Stereo pairs are sorted according to the intersection angle, incidence angle and proximity of acquisition date.
- **SIFT order:** The number of pairwise matches can be interpreted as a measure of the shared visual content between two images. This order sorts all possible pairs in decreasing number of SIFT matches, therefore prioritizing pairs with a higher overlap of visual content.

In all experiments, the best (i.e. the first) 50 pairs according to each selection criterion were employed to reconstruct the AOIs. Due to the high computational cost, the oracle order was only computed for the IARPA dataset.

3.2. Results and discussion

Table 1 reflects some of the concepts anticipated in Section 2. First of all, we can verify that naif bundle adjustment (single run, classic L^2 norm for reprojection errors) produces worse DSMs compared to the rest, even when optimal stereo pairs are used, underlining the need of strategies to handle outliers in the feature tracks.

Table 1 also supports the assumption that adjusting a 2D translation at image level or the 3D rotation of a satellite camera is almost equivalent, with the results being extremely similar in all experiments involving these strategies.

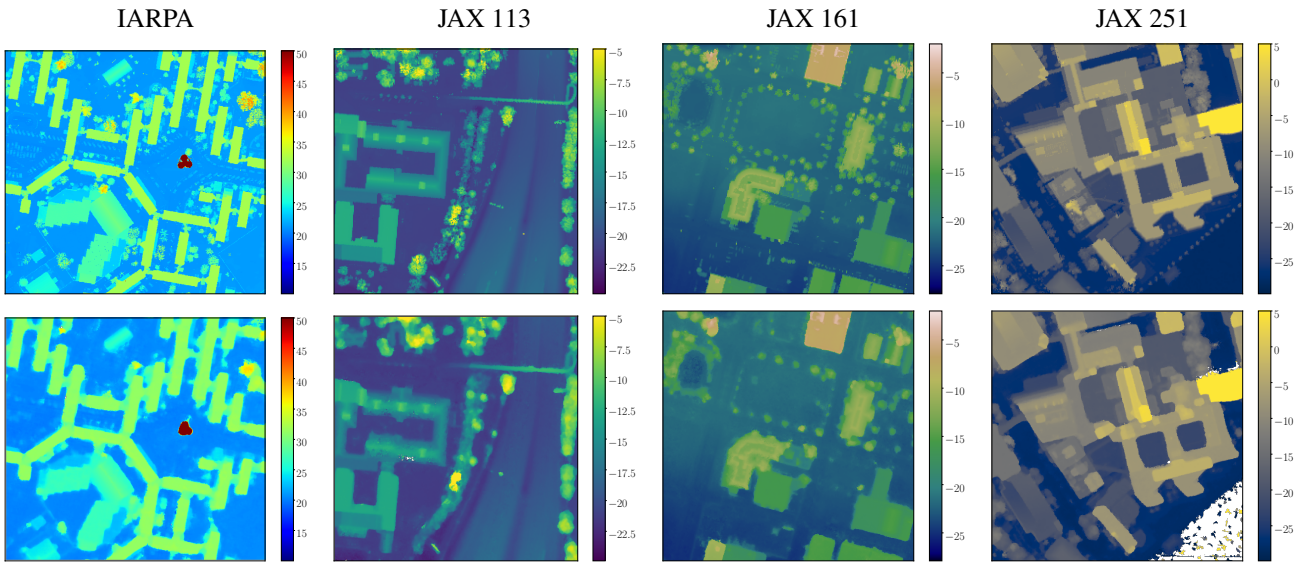

Figure 4: Lidar acquired DSMs versus 3D reconstruction from multi-date satellite images. The ground truth DSMs from the IARPA and GRSS datasets are shown on the top row. For each AOI, the reconstructed DSM that obtained the highest completeness score is displayed below. Colorbars on the right of the DSMs assign a color to each height, expressed in meters. White points represent unknown values. The color palettes were chosen to optimize the view of scene details.

It follows that complex methods to initialize the sparse point cloud input to the bundle adjustment, as Algorithm 3, can be replaced by simpler and more efficient methods such as the one employed in Section 2.3.

In all experiments, the overall RMSE of correlation based DSM alignment is slightly smaller. This bias seems natural considering that the maximum NCC is minimizing the difference in location between all points of the DSMs; whereas bundle adjustment strategies only use a reduced amount of keypoints (i.e. feature tracks) to register them.

In any case, the main idea that seems to stand out from the majority of results is that both geometry and image based solutions are valid and competitive solutions to the problem. Most of the experiments produced DSMs with completeness score above 65%. However, in certain scenarios, some of the pipelines experimented a loss of accuracy.

Failure prone cases for image based corrections. For the IARPA AOI, the correlation based DSM alignment clearly outperformed bundle adjustment methods when using the number of SIFT matches to select input stereo pairs instead of the oracle or the heuristic orders.

The loss of accuracy of bundle adjustment methods, illustrated in Figure 5, can be explained by looking at the connectivity graph of the images according to the number of SIFT matches: it turns out that there is a group of 5 nodes weakly connected to the rest. The 5 nodes correspond to images taken from a similar viewpoint, with a large incidence angle. Consequently, they have very strong intra-similarity but less resemblance to the rest. If we only display edges accounting for more than 40 matches, the 5 nodes (in red) are disconnected from the others (see Figure 6, IARPA). The heuristic order does not use these views because they are too tilted. Still, red nodes exhibit a large amount of matches between them, so several pairs from the set are selected by the SIFT order. What happens then is that bundle adjustment may end up putting all white nodes into a common frame of reference, to the extent possible, while red nodes are adjusted to a frame that fits better their particular set of observations. This can be verified by exploration of the DSMs obtained from pairs of white and red nodes.

Oppositely to bundle adjustment, the correlation based DSM alignment offers higher robustness to this scenario, since it was conceived to deal with non registered DSMs.

Based on the preceding, we can state that bundle adjustment algorithms require not only quality feature tracks to work properly, but in addition such tracks should connect the graph of input images in a consistent manner. Otherwise, it is better to avoid incorporating stereo pairs from disconnected sets, as it happens when the oracle order is used (i.e. no pair takes both views from the set of red nodes).

In contrast to the IARPA case, the heuristic and SIFT orders yield similar results for the JAX AOIs. The connectivity graphs of the Jacksonville images are much more consistent (see Figure 6, JAX 113). Accordingly, the use of SIFT matches as a selection criterion seems appropriate. For JAX AOIs, more than 20 of the pairs selected by the SIFT order are also chosen by the heuristic order; for IARPA, only 2 pairs coincide in both orders.

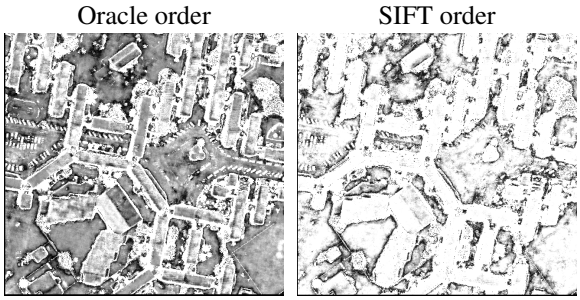

Figure 5: Reconstruction error of the DSM obtained with bundle adjustment of camera rotations (IARPA AOI), using the oracle and the SIFT orders. Brighter values account for larger errors. Errors above 1 m are clipped and correspond to white pixels.

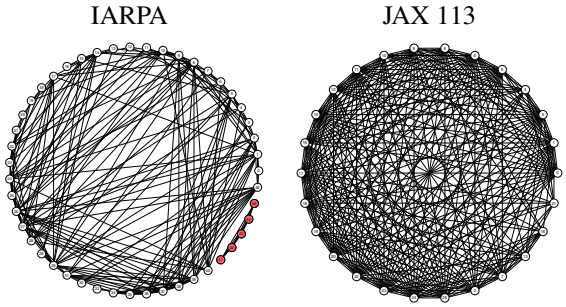

Figure 6: Connectivity graphs for the images of the IARPA and JAX 113 AOIs (edges link images with more than 40 matches).

Failure prone cases for geometry based correction. In Table 1 we can see that some results obtained by DSM alignment are slightly worse than the bundle adjustment counterparts (e.g. JAX 251 and IARPA with heuristic order). Since the DSMs being fused are the same, these errors are due to the alignment itself. They can be attributed to the fact that correlation based alignment is sensitive to incomplete geometry, especially if the holes in the DSMs are relevant with respect to the size of the AOI. Even in the absence of major radiometric changes, incomplete (i.e. unknown) areas in the DSM may be caused by occlusions, which are common in areas with many structures, as is the case in the concerned sites; or by water bodies, as in the lower right corner of JAX 251.

Deficiencies not due to pointing error. There is an amount of error in the output DSMs that has nothing to do with RPC inaccuracies. Even when these are properly corrected, errors at the edges of buildings or in vegetation areas are typically larger than 1 m (see Figure 5, oracle order). This is not surprising since in these areas it is necessary to choose between two extremely different modes: a 3D point belongs either to the floor or to a rooftop/tree. Thus, the values of the DSMs can be expected to have higher variance there and using a median filter to fuse them is insufficient. Further discussion on the multi-modality of DSM heights can be found in [11].

4. Conclusion

We reviewed and compared 3 MVS pipelines for automatic 3D reconstruction from multi-date satellite images, each of them using a different method to correct the pointing error of satellite RPCs: (1) geometry correlation to align independent DSMs, (2) bundle adjustment to optimize 2D correction offsets in the image domain (3) bundle adjustment to optimize the 3D rotation of satellite cameras.

All the approaches proved to be valid and competitive solutions to the problem. Overall, they achieved very similar evaluation metrics, but in some cases differences in performance were revealed. The geometry based strategy emerged as a more robust solution when stereo pairs from weakly connected subsets of images (in terms of feature matching) are used as input. Oppositely, image based solutions relying on bundle adjustment proved to be more robust in lack of geometry, either because the area of interest is too small, or because the site can only be reconstructed partially. The presence of water or large occlusions can be possible reasons for the latter case.

It seems clear that there is room for improvement regarding the relative geolocation correction of satellite RPCs. This work unveiled some clues about which scenarios seem to be more prone to failure for each method, but further research needs to be carried out on a larger scale to confirm the presented findings and draw sound conclusions.

Future work. The conducted experiments highlighted that sub-optimal selections of stereo pairs make it harder for geolocation correction strategies to success. It was shown that considering the distribution of pairwise matches can be a valuable additional source of information to discard inconvenient views or pairs. This insight could complement the heuristics used in the literature to select input pairs, mainly focused on the images metadata. The fact that the metadata affects the entire satellite images and is not particular to specific areas of interest argues in favour of this idea.

Adjusting a set of correction rotations directly composed with the original RPCs, without relying on the affine camera approximation, may also be useful to handle larger AOIs.

Last but not least, the investigated algorithms are not incompatible. Probably an intermediate two-step pipeline, using bundle adjustment followed by an ideally redundant DSM alignment would offer even higher robustness, since both image data and geometry would be exploited.

Acknowledgements. Work partly financed by Office of Naval research grant N00014-17-1-2552, DGA Astrid project « filmer la Terre » n^o ANR-17-ASTR-0013-01.

References

[1] M. Bosch, K. Foster, G. Christie, S. Wang, G. D. Hager, and M. Brown. Semantic Stereo for Incidental Satellite Images. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1524–1532. IEEE, jan 2019. 2

[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. 2, 3, 6

[3] CMLA and CNES. Satellite Stereo Pipeline S2P source code, 2019. 3

[4] P. D'Angelo and G. Kuschk. Dense multi-view stereo from satellite imagery. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2012. 2

[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, jul 2012. 2

[6] 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. 2, 3

[7] C. de Franchis, E. Meinhardt-Llopis, J. Michel, J.-M. 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. 2, 3

[8] G. Dial and J. Grodecki. RPC replacement camera models. In *ASPRS 2005 Annual Conference*, pages 1–9, 2005. 1

[9] L. Duan and F. Lafarge. Towards large-scale city reconstruction from satellites. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016. 2

[10] G. Facciolo, C. de Franchis, and E. Meinhardt. MGM: A Significantly More Global Matching for Stereovision. In *Procedings of the British Machine Vision Conference 2015*, pages 90.1–90.12. British Machine Vision Association, 2015. 3

[11] G. Facciolo, C. De Franchis, and E. Meinhardt-Llopis. Automatic 3D reconstruction from multi-date satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 57–66, 2017. 2, 3, 6, 8

[12] W. Förstner and E. Gülch. A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centres of Circular Features. In *Proceedings of ISPRS Intercommission Conference on Fast Processing of Photogrammetric Data*, pages 281–305, 1987. 2, 3

[13] K. Gong and D. Fritsch. Point cloud and digital surface model generation from high resolution multiple view stereo satellite imagery. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 42(2), 2018. 2, 3, 6

[14] J. Grodecki and G. Dial. Block adjustment of high-resolution satellite images described by rational polynomials. *Photogrammetric Engineering & Remote Sensing*, 69(1):59–68, 2003. 2, 3, 5

[15] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2003. 4

[16] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2008. 2

[17] B. Le Saux, N. Yokoya, R. Hansch, M. Brown, and G. Hager. 2019 Data Fusion Contest [Technical Committees]. *IEEE Geoscience and Remote Sensing Magazine*, 7(1):103–105, mar 2019. 2

[18] M. J. Leotta, C. Long, B. Jacquet, M. Zins, D. Lipsa, J. Shan, B. Xu, Z. Li, X. Zhang, S.-F. Chang, M. Purri, J. Xue, and K. Dana. Urban Semantic 3D Reconstruction from Multiview Satellite Imagery. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 2

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 3

[20] P. Moulon and P. Monasse. Unordered feature tracking made fast and easy. In *CVMP 2012*, page 1, 2012. 3

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

[22] O. C. Ozcanli, Y. Dong, J. L. Mundy, H. Webb, R. Hammoud, and T. Victor. Automatic geo-location correction of satellite imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 307–314, 2014. 1, 2, 3, 4

[23] E. Runpik, M. Pierrot-Deseilligny, and A. Delorme. 3D reconstruction from multi-view VHR-satellite images in MicMac. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2018. 2

[24] 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. 2, 3

[25] 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. Springer, 2000. 2