L1B$^+$: A PERFECT SENSOR LOCALIZATION MODEL FOR SIMPLE SATELLITE STEREO RECONSTRUCTION FROM PUSH-FRAME IMAGE STRIPS

Roger Marí[1,*] , Thibaud Ehret[1] , Jérémy Anger[1,2] , Carlo de Franchis[1,2] , Gabriele Facciolo[1]

[1] Université Paris-Saclay, CNRS, ENS Paris-Saclay, Centre Borelli, 91190, Gif-sur-Yvette, France [2] Kayrros SAS

KEY WORDS: Rational Polynomial Coefficients, RPC model, push-frame, SkySat, satellite stereo

ABSTRACT:

We propose a novel method to generate a single image product from a multi-image strip acquired by a push-frame satellite imaging system. The images of the push-frame strips are combined into a large scale mosaic simulating a perfect sensor geometry. The local camera models of the input images are leveraged to produce a new localization model that covers the output mosaic entirely. Among other applications, this simplifies the task of stereo reconstruction enormously: instead of treating multiple stereo pairs of small images, it is possible to reconstruct the entire area covered by the push-frame acquisition using a single pair of mosaics incorporating all the images. We test our method using strips of SkySat L1B scenes and denote the output images as L1B$^+$. To evaluate the quality of the L1B$^+$ images and their localization models, the stereo reconstructions obtained with L1B$^+$ are compared with those obtained with L1B and with a lidar reference model.

1. INTRODUCTION

Push-frame satellite imaging systems acquire a continuous strip of small and partially overlapping images as the satellite moves (Aati and Avouac, 2020, Planet, 2021). Small satellite (SmallSat) constellations, such as SkySat and PlanetScope from Planet or Aleph-1 from Satellogic, use a push-frame imaging mode to cover large areas of interest, beyond the limited footprint of their telescope. The resulting image collections are a product of great interest: small satellites can orbit at low altitudes and offer very high resolution, in direct competition with well-established large satellite providers. While the latter are capable of covering large areas with a single shot, the derived products are also more expensive. SmallSat constellations can also afford many more units, resulting in shorter revisit times (Sandau et al., 2010). However, despite their economic and technical appeal, the highly fragmented nature of push-frame acquisitions may discourage their use for some applications targeting areas of interest of several km^2 (Xue et al., 2008).

We propose a novel method to exploit push-frame image collections from SmallSat acquisitions, which circumvents the difficulties of working with such fragmented data. The method combines the multiple images of a strip acquired by a push-frame system into a single mosaic image, as if it had been acquired by an instrument with perfect sensor geometry (Figure 1). The local camera models of the input images are leveraged to produce a new localization model that covers the output mosaic entirely. The resulting image and localization model can be treated as new product, which we denote L1B$^+$.

Our contributions are:

- A method capable of assembling the frames and camera models of strips of partially overlapping satellite images.
- An evaluation of the method based on an application of major interest: stereo reconstruction. We validate the method using strips of SkySat L1B scenes. The 3D models obtained with the L1B$^+$ products are compared with those obtained with L1B and with a lidar reference model.

* Corresponding author (roger.mari@ens-paris-saclay.fr)

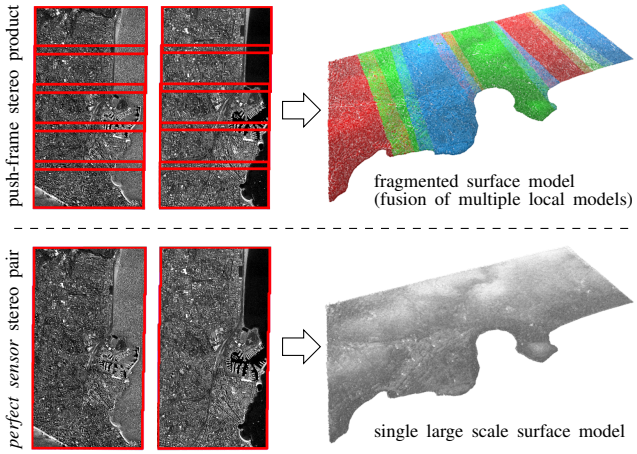

Figure 1. We combine the multiple images of each strip acquired by the push-frame system to produce large scale mosaics. The resulting *perfect sensor* images and their localization models can be used to greatly simplify the task of large scale stereo reconstruction from push-frame satellite imagery.

2. RELATED WORK

SmallSat push-frame imagery is used in a wide range of remote sensing applications, including topography extraction (Aati and Avouac, 2020, Bhushan et al., 2021, d'Angelo and Reinartz, 2021), super-resolution products (Nguyen et al., 2021, Anger et al., 2020) and various tasks demanding short revisit times, such as monitoring of natural phenomena (Cannistra et al., 2021) or commercial assets (Marí et al., 2021b, d'Autume et al., 2020). This paper focuses on the task of 3D reconstruction, but the proposed method can be beneficial for any of these applications.

Topography extraction from high resolution satellite images is typically performed using stereo pipelines (Beyer et al., 2018, de Franchis et al., 2014a), capable of producing highly accurate photogrammetric digital surface models (DSMs). These pipelines take as input one pair of images and their geometric camera models, represented with the rational polynomial coef-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-1-2022-137-2022 | © Author(s) 2022. CC BY 4.0 License.

ficients (RPC) model. The RPC model represents the image acquisition system by means of a pair of rational polynomial functions that approximate the mapping from 3D space points to 2D image pixels: $\mathcal{P} : \mathbb{R}^3 \rightarrow \mathbb{R}^2$, the projection function; and its inverse $\mathcal{L} : \mathbb{R}^2 \times \mathbb{R} \rightarrow \mathbb{R}^3$, the localization function. The RPC camera models allow to represent unconventional geometric models such as that of push-broom scanners. Push-broom cameras are not projective as the optical center changes from line to line. However, for satellite images they can be approximated locally by affine models (de Franchis et al., 2014b). This is exploited by stereo pipelines (de Franchis et al., 2014a, Beyer et al., 2018), which cut the images into small tiles that can be stereo-rectified and processed independently. This insight motivates our merged camera model for push-frame images and its usefulness in the stereo context.

Image collections acquired by push-frame systems cannot be directly plugged into the state-of-the-art 3D reconstruction pipelines, as the task becomes a multi-view problem in which each image will contribute only to some part of the final reconstruction. The majority of solutions that can be found in the literature address the reconstruction of an area of interest observed by multiple overlapping satellite images as a two-stage multi-view stereo (MVS) problem (Facciolo et al., 2017, Gong and Fritsch, 2018, Bhushan et al., 2021). In the first stage, a series of local models are computed, resulting from independent executions of a stereo reconstruction pipeline using different input pairs. In the second stage, the local models are fused to obtain the complete reconstruction of the area. This two-stage approach requires significant pre- and post-processing work.

The pre-processing tasks of satellite MVS require the selection of an optimal set of stereo pairs, to minimize the number of local models to be reconstructed and all the derived workload. In addition to the amount of geographic footprint overlap, the incidence angles of the cameras, the angle between both views and the acquisition dates have proven to be important to select suitable stereo pairs (Facciolo et al., 2017, Gong and Fritsch, 2018, Marí et al., 2019, Gómez et al., 2022). Another classic pre-processing step is the correction of camera positions and orientations, in order to make the multiple views geometrically consistent. Bundle adjustment algorithms have so far stood out as the best practice to correct the inaccuracies of the RPC models (Triggs et al., 1999, Grodecki and Dial, 2003, Marí et al., 2021a). The corrected RPC models produced by the bundle adjustment methods minimize the reprojection error of a set of reference points observed across the different images. The differential advantage of adding a camera correction step prior to stereo reconstruction is that local models computed from independent pairs are, in principle, natively registered in the object space.

The post-processing tasks of satellite MVS revolve around the fusion step. Assuming that all local models are aligned, the usual procedure is to discretize the area of interest and apply a mean or median filter on each 2D cell to establish the final altitude. However, the result of the fusion is usually largely improvable and several works have introduced heuristics (Facciolo et al., 2017, Qin, 2017) or deep learning refinement strategies (Stucker and Schindler, 2020) to minimize artifacts caused by outliers, vegetation or inexact geometry registration. In awareness of the complexity of the process, the NASA Ames pipeline (Beyer et al., 2018) has been expanded with the addition of optional processing tools including bundle adjustment and point cloud registration algorithms.

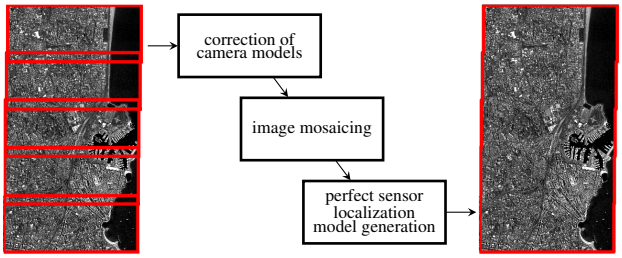

Figure 2. Diagram of the presented methodology.

In the context of 3D reconstruction from push-frame images, we eliminate much of the previous pre- and post-processing work by addressing the root of the problem: the format of the data. The input images are combined to simplify the MVS problem into an ideal scenario with a single input pair. The proposed perfect sensor geometry model for push-frame strips takes inspiration from large push-broom satellites such as in the Pléiades-HR constellation. Due to the complexity of the focal plane, the Pléiades-HR raw products should be considered as 25 sub-products with their local geometrical models (Baillarin et al., 2010). However, for the sake of usability, the final images emulate the geometry of an ideal push-broom linear array. The Pléiades-HR perfect sensor geometry models are derived from the raw image, the rigorous sensor model of the satellite and a coarse elevation model of the area (Baillarin et al., 2010). Another existing tool that follows a similar philosophy is the `dg_mosaic` from the Ames pipeline, which can mosaic multiple subscenes derived from the same parent Maxar push-broom product and create a new RPC model for it (Shean et al., 2016). In this work we address a more generic problem, as the input scenes do not originally belong to a common image.

3. METHODOLOGY

Given a push-frame strip S_1 of N small footprint images (also known as *scenes*), we propose a method to generate a single and equivalent large footprint image, denoted S_1^+, along with its perfect sensor localization model. In this work, we assume that the input images are radiometrically calibrated and cloud-free. As shown in Figure 2, the method consists of three main steps, which are detailed in the following subsections.

3.1 Correction of camera models

First of all, it is necessary to ensure that the local RPC camera models of all the scenes in S_1 are geometrically consistent. That is, 3D points in object space project to corresponding points in each image. Enforcing the consistency will ease the subsequent mosaicing step and it will be fundamental to produce the perfect sensor localization model of the mosaic S_1^+. This is traditionally achieved by bundle adjustment in a multi-view setting. However, since the baseline between the consecutive cameras of the same strip is too small for 3D vision purposes, some additional information is needed. To this end, we employ a secondary strip S_2 covering the same area of interest observed by S_1 but from another point of view. SkySat stereo or tri-stereo products can provide the secondary strip S_2. Given the collection of $2{\cdot}N$ scenes in S_1 and S_2, the bundle adjustment methodology proposed in (Marí et al., 2021a) is applied to perform a relative correction of their local RPC models. Each corrected RPC model results from composing the original projection function with a corrective rotation transform around the estimated camera center. The corrective rotation compensates for the

main source of inaccuracies, which is the inexact knowledge of the attitude angles. The reference points used by the bundle adjustment are automatically generated from correspondences of SIFT keypoints (Lowe, 2004).

Note that the camera correction step does not restrict the scope of the presented method to stereo push-frame acquisitions. It is also possible to correct the camera models of S_1 without a secondary strip if a basemap with reference DSM or a set of ground control points (GCPs) are available. GCPs are points with known 3D coordinates whose position in the images is also available (Grodecki and Dial, 2003).

3.2 Image mosaicing

Once the camera models are corrected, each scene I_i in S_1 is warped into a common image space using a 2D projective transform H_i. A rough estimate of H_i is first obtained by establishing 2D point correspondences between the i-th scene and the central scene of the strip, i.e. the $\frac{N}{2}$-th scene. Instead of using classic feature matching techniques, we take advantage of the corrected RPC models computed in the previous step to initialize H_i (Section 3.2.1). Then, each H_i is refined into $\hat{H}_i$ using an image registration method (Baker and Matthews, 2001, Briand et al., 2018) (Section 3.2.2). Lastly, the output mosaic S_1^+ is obtained by averaging all the warped scenes as determined by the corresponding $\hat{H}_i$. A high-order Spline interpolation is used to perform the warping (we use order 5).

3.2.1 Image warping initialization For each scene I_i, a regular grid of 10×10 2D points is localized in the 3D space using the i-th RPC localization function $\mathcal{L}_i$ and the average altitude of the area, h_{avg}. The average altitude h_{avg} of the area may be only an approximation, for instance a rough estimation can be obtained using the SRTM data (Farr et al., 2007). Each grid point is then reprojected into the image space of the central scene of the strip, using the corresponding RPC projection function $\mathcal{P}_{\frac{N}{2}}$. The reprojection results in a set of 2D correspondences between a point $\boldsymbol{x}$ from each scene I_i and its homologous $\hat{\boldsymbol{x}}$, located in the mosaic image space, where the central scene of the strip remains in the center. Equation 1 summarizes the previous procedure:

$$\hat{\boldsymbol{x}} = \mathcal{P}_{\frac{N}{2}}(\mathcal{L}_i(\boldsymbol{x}, h_{\text{avg}})) \quad \text{where } i \in [1, N]. \tag{1}$$

The correspondences $\boldsymbol{x} \leftrightarrow \hat{\boldsymbol{x}}$ are then used to fit (using a classic DLT algorithm (Hartley and Zisserman, 2004)) homographic transformations $H_i \in \mathbb{R}^{3\times3}$ such that $\hat{\boldsymbol{x}} = H_i\boldsymbol{x}$.

Note that in the case of the strip scenes that do not overlap with the central scene, we simply localize and reproject recursively along the neighboring frames, in the direction of the central scene, until we reach the latter.

3.2.2 Image warping refinement The correspondences used to compute the transform H_i are inaccurate because the h_{avg} value used for the reprojection is not the exact altitude of the points seen in the image. However, the estimated transformations are useful to initialize an image-based registration method. The inverse compositional algorithm (ICA) is used to refine the coefficients of H_i, so that the warps of consecutive frames are precisely aligned. This assures the pixel consistency of the aligned scenes before merging them into the S_1^+ mosaic. The benefits of ICA to register push-frame satellite acquisitions has been previously studied by (Anger et al., 2020, Briand et al., 2018).

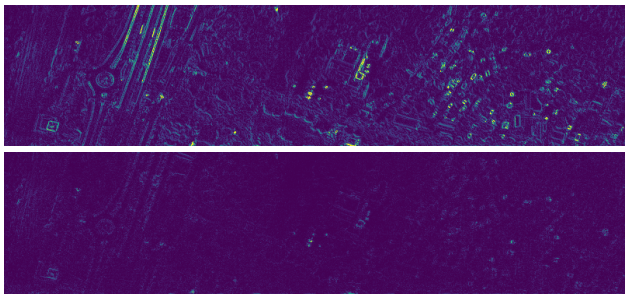

Figure 3. Residual difference in the overlap region after the alignment of two consecutive SkySat L1B scenes of the same push-frame strip. Top to bottom: using the initial transform H_i, using the refined transform $\hat{H}_i$. The same scaling and colormap have been used for both residuals.

Consider two scenes I and I' of S_1 such that I' has to be aligned onto I to construct the mosaic S_1^+. Let H and H' be the initial homographies associated to I and I' respectively, which are computed as described in Section 3.2.1. Using ICA, we first refine $H' \circ H^{-1}$ such that the warped version of I' using this transformation is perfectly aligned onto I. This defines the refined homography R. Using the refined transform, we define the relative correction factor C such that $R = H' \circ C$. Figure 3 shows the residual difference after alignment with and without the refinement step. Observe that the refined transformation achieves a much better alignment. In this example, the RMSE without refinement is 19.37 and with refinement 5.71, thus confirming the visual result.

We then define by recurrence the set of refined transforms $\hat{H}_i = H_i \circ C_i$, where C_i corresponds to the composition of all necessary correcting factors from the reference image (the central scene) to the i-th image in S_1.

3.3 Perfect sensor geometry localization model

After completion of the mosaic S_1^+, the corrected camera models of the scenes that form the mosaic can be used to produce a perfect sensor localization model that follows the RPC standard and is valid throughout the entire S_1^+. The output camera model, denoted $\text{RPC}_{S_1^+}^+$, is generated by Algorithm 1.

The main idea of Algorithm 1 is to draw a regular grid of 2D points covering S_1^+, which is localized at different heights in the 3D space. By default, we use $(N\cdot M)\times(3\cdot M)$ points, where N is the number of scenes and $M = 10$. Given a 2D point $\hat{\boldsymbol{x}} \in S_1^+$, the corresponding 3D point $\hat{\boldsymbol{X}}$ at height h is obtained as

$$\hat{\boldsymbol{X}} = \mathcal{L}_i(\hat{H}_i^{-1}\hat{\boldsymbol{x}}, h), \tag{2}$$

where $\hat{H}_i^{-1}$ is the inverse warping transform that transforms $\hat{\boldsymbol{x}}$ to its original small scene space, and $\mathcal{L}_i$ is the localization function of that scene. We set the range of altitudes $[h_{\text{min}}, h_{\text{max}}]$ by taking the maximum and minimum altitudes of the reference points used by the bundle adjustment (Marí et al., 2021a) and adding an extra margin of ±100 meters.

Thanks to the corrected camera models, the different local RPC functions are highly consistent in the object space. This implies that the localization of the grid will result in a reasonably regular point cloud in the object space, without major discontinuities due to RPC inaccuracies. Furthermore, when a point $\hat{\boldsymbol{x}}$ is seen in two overlapping scenes I_i and I_j, it can be localized in object space, with (2), using either $\mathcal{L}_i$ or $\mathcal{L}_j$. Since the scenes

Algorithm 1: *Perfect sensor localization model generation*

Input : 1 strip S_1 of N partially overlapping scenes I_i,
with their local RPC$_i$ camera models ($i \in [1, N]$).
S_1^+, the mosaic image of S_1,
and the matrices $\hat{H}_i$ that warp scene I_i onto S_1^+
Output: 1 RPC$_{S_1}^+$ camera model covering the entire S_1^+

1. Build G_{2D}, in the image space:
   Draw a regular grid of $(N{\cdot}M){\times}(3{\cdot}M)$ 2D points on S_1^+
2. Build G_{3D}, in the object space:
   Mark all the points of G_{2D} as non-visited
   for h **in** $3{\cdot}M$ uniformly spaced altitudes $\in [h_{\min}, h_{\max}]$
     for each scene I_i **in** S_1
       Localize at h the non-visited points of G_{2D} seen in scene I_i, using (2), and mark them as visited
3. Use the $(N{\cdot}M) \times (3{\cdot}M) \times (3{\cdot}M)$ correspondences $G_{2D} \leftrightarrow G_{3D}$ as input for (Akiki et al., 2021) to fit RPC$_{S_1}^+$

are registered and the RPC models adjusted, both choices will yield essentially the same 3D points for a range of altitudes centered around the surface. However, for points far from the surface we should start to observe a parallax due to the fact that the scenes are acquired from different positions along the orbit. This parallax can be quantified. Assuming the SkySat parameters (baseline between views ∼1.5 km and altitude ∼500 km), in order to observe parallax of 1 pixel (assuming a resolution of 0.6 m/pixel) an elevation change of about 200 m $\approx \frac{500}{1.5} \cdot 0.6$ m should be present in the scene. This points to a limitation of the present method for images containing very large elevation changes. However, in our experiments (including the mountainous site of the Morenci mine, see Section 4) we did not observe any artifacts due to parallax.

The previous procedure results into a set of 2D-to-3D point correspondences between S_1^+ and the object space, thus the RPC fitting algorithm from (Akiki et al., 2021) can be applied to produce the final RPC$_{S_1}^+$ model.

Using the 2D-to-3D correspondences generated earlier with (2) we can define the fitting errors e associated to the RPC$_{S_1}^+$ model as the reprojection distances:

$$e = \|\hat{\boldsymbol{x}} - \mathcal{P}_{S_1^+}(\hat{\boldsymbol{X}})\|_2, \tag{3}$$

where e is in pixel units and $\mathcal{P}_{S_1^+}$ is the projection function of the perfect sensor localization model. Figure 4(b) illustrates the usual distribution of errors e across the 3D points used to fit RPC$_{S_1}^+$. We observe that the error is small in the proximity of the surface, which can be inferred from the reference keypoints (seen as blue dots) used in the bundle adjustment (Section 3.1). This is reasonable as the surface points are registered in the merged product. Larger errors are observed approaching the altitude extrema of the volume, but only in bands that correspond to the overlap of two consecutive scenes. We attribute this to the fact that there is no guarantee that the RPCs of neighboring scenes are geometrically consistent away from the registered surface points. Inside the convex hull that contains all the reference points, e reaches average values ∼0.2 pixels, of the same order as the average reprojection error of the bundle adjustment (Marí et al., 2021a). Note that the 3D points used by the bundle adjustment highlight the part of the volume where the surface observed by S_1^+ is located.

4. EXPERIMENTS

4.1 Data

We applied our method to two SkySat L1B stereo acquisitions, each providing two multi-image strips of the same area, with a time difference of a few seconds. One acquisition covers part of the city of Antibes (France) and the other covers the Morenci mine (United States). The two landscapes are very different: urbanized and flat terrain in the case of Antibes, as opposed to the mountainous and bare terrain of the mine.

SkySat L1B scenes have a nadir resolution of 0.58-0.86 m/pixel and a total size of 1349×3199 pixels. Each scene is delivered with an RPC camera model. The geometric accuracy of the L1B RPC models is of 30-50 m, with SkySats orbiting at altitudes of 400-500 km (Planet, 2021). Our proposed L1B$^+$ mosaics extend the footprint of the original L1B images and incorporate a consistent RPC model. In this paper, we present experiments using N = 3 and N = 5 scenes per strip but the method can generalize to strips with more scenes. Using SkySat L1B scenes, we observed no deformation or misalignment in the output mosaics with strips with a number of scenes up to N = 13.

The SkySat acquisition platform has three staggered sensors, resulting in the push-frame system simultaneously acquiring three multi-image strips. Note that in this work we assemble images from only one of the sensors at a time.

4.2 Stereo reconstruction based evaluation

To validate the quality of the L1B$^+$ images and their camera models, we evaluate them in the context of stereo reconstruction from two push-frame strips S_1 and S_2. For this purpose, we used the open-source satellite stereo pipeline S2P[1] (de Franchis et al., 2014a), to reconstruct the areas of Antibes and the Morenci mine covered by the SkySat acquisitions, both using the original L1B scenes and the L1B$^+$ mosaics, i.e. S_1^+ and S_2^+.

As explained in Section 2, for the case of the L1B scenes the 3D reconstruction is a multi-view problem. We use the MVS methodology described in (Marí et al., 2021b) to solve it. Following the selection of P suitable pairs of scenes, S2P is employed to reconstruct P independent local DSMs, using the corrected RPCs of the L1B scenes (Section 3.1). The P local models are lastly fused by taking the mean altitude at each cell of the DSM. In the conducted experiments, P = 5 for N = 3, while P = 9 for N = 5.

In contrast with the above, the L1B$^+$ products allow to reconstruct each area with a single execution of S2P, using as input the two perfect sensor images and their localization models.

4.3 Discussion

Table 1 lists the mean absolute error (MAE) between the DSMs obtained with the L1B$^+$ and L1B products. In addition, we computed the MAE of each model with respect to a ground-truth (GT) lidar model covering a subregion of the observed areas. Since the acquisition dates of the lidar and the SkySat images are not coincident, we manually annotated the parts of the surface models that are expected to be coherent.

[1] `https://github.com/centreborelli/s2p`

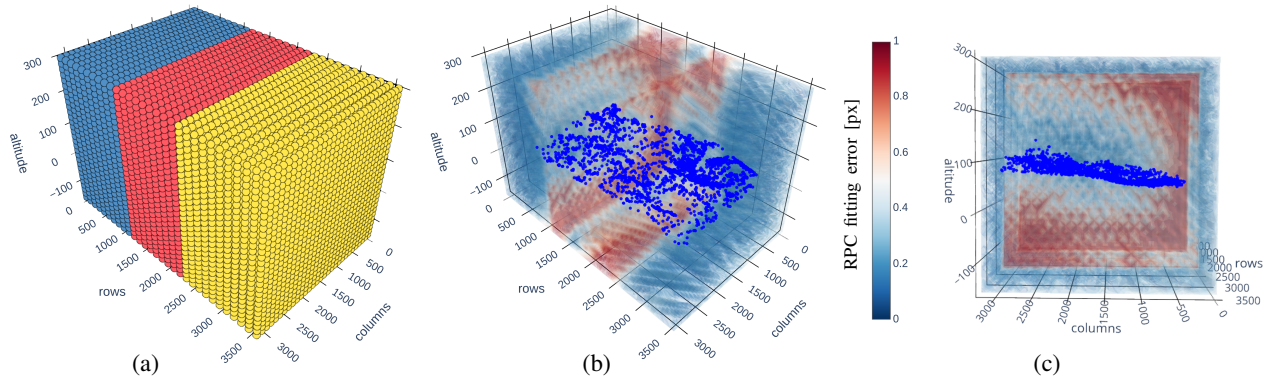

Figure 4. (a) 3D grid used for RPC fitting, for a strip of 3 scenes. The z coordinate is the altitude of the points in meters, while x and y correspond to their projection in the image plane. Point colors depend on the scene of the strip that was used to localize each 3D point. (b) Error of the perfect sensor localization model across the 3D grid, in pixel units. The reference points used by the bundle adjustment (Marí et al., 2021a) in the prior correction of the local RPCs are shown in blue. (c) Front view of the error distribution.

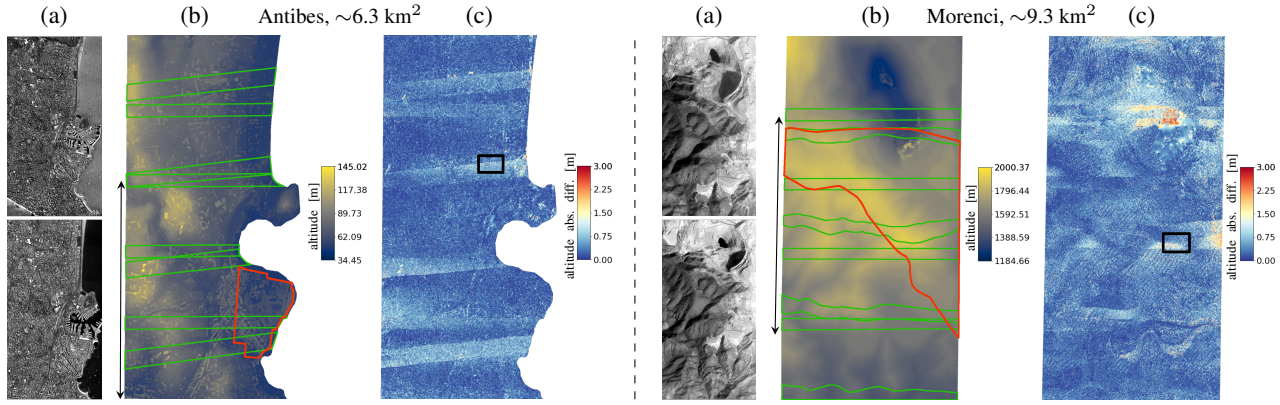

Figure 5. Qualitative results for SkySat stereo reconstruction, using 5 scenes from the input strips, i.e. $N = 5$. Left to right: (a) L1B$^+$ images. (b) L1B$^+$ derived DSMs. The size of the reconstructed area is indicated in square kilometers. The red boundary delimits the region for which a lidar model is available. The green boundaries delimit the regions of overlap between two local models used to derive the equivalent L1B DSM. The double-headed arrow indicates the length covered in the experiments with $N = 3$. (c) Absolute difference between the L1B$^+$ and L1B DSMs. The black rectangular outline delimits a subregion of interest inspected in Figure 6.

		MAE [m]		
SkySat strip IDs	N	L1B$^+$-L1B	L1B$^+$-GT	L1B-GT
s107_20210705T131230Z s107_20210705T131300Z (Antibes)	3	0.325	1.214	1.270
	5	0.331	1.204	1.248
s4_20190127T175119Z s4_20190127T175154Z (Morenci)	3	0.391	0.781	0.854
	5	0.395	0.745	0.780

Table 1. Quantitative results of the stereo reconstruction based evaluation, using 3 or 5 scenes from the input strips. Left to right: MAE, in meters, between the L1B$^+$ and L1B derived DSMs, and MAE of each DSM with respect to a GT lidar model.

Figure 5 shows the L1B$^+$ images for $N = 5$, the resulting L1B$^+$ DSMs and the absolute difference with respect to the equivalent L1B DSMs. In Figure 5(c), we can see that the absolute difference between L1B and L1B$^+$ DSMs is below 0.3 m in the majority of the surface points (the average corresponds to the L1B$^+$-L1B column of Table 1). However, there are parts of the area where this difference increases and approaches values close to 1 m. These traces are a consequence of the fusion of local models that is needed to generate the L1B DSM. In fact, the traces coincide in great measure with the green boundaries in Figure 5(b), which indicate the areas of overlap between the local models used to produce the L1B DSM. The local models do not match perfectly, because the correction of the camera models registers their geometries with an accuracy < 1 m, but a residual remains (Marí et al., 2021a). Consequently, the average altitude retained by the fusion process is subject to a certain degree of bias, especially in these overlapping zones. In the case of Antibes, the areas showing the largest differences follow a pattern of horizontal stripes, because the local geometries consist of overlapping planes (flat terrain), which are stacked along the vertical axis. In the case of the Morenci mine the pattern is more irregular, with curves caused by non-exactly coincident mountain shapes and peaks due to the presence of outliers near the open pit (close to the upper left corner).

In Figure 6, we selected two subregions where the altitude differences between the L1B$^+$ and the L1B DSMs exhibit a strong increase. In accordance with the above observations, we can see that such differences are indeed caused by biases in the L1B DSM. By using L1B$^+$ products we eliminate the cause of such biases, i.e. the need to register and merge any local models, so the discontinuities disappear in a natural way.

The last two columns of Table 1 indicate that the L1B$^+$ DSMs improve the accuracy of the L1B ones, as they exhibit smaller differences with respect to the lidar. The MAE values obtained with L1B$^+$ are quite stable too, regardless of whether 3 or 5

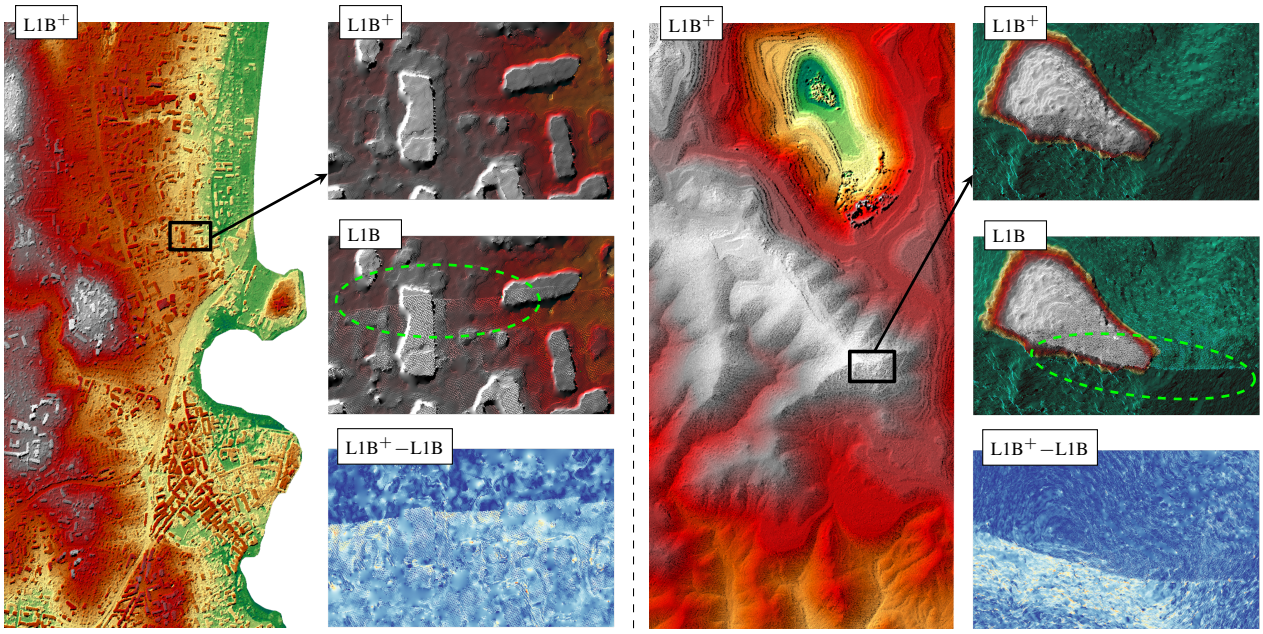

Figure 6. Detailed view of the L1B$^+$ DSMs and a subregion of interest. The inspection of the subregions shows that, in absence of outliers, the largest differences between L1B$^+$ and L1B DSMs coincide with small altitude discontinuities (circled in green) in the L1B model. The colormap is different in the subregion images with respect to the complete DSM to improve the contrast between local altitude values. The L1B$^+$ −L1B images represent the absolute difference between altitudes, with the same colormap of Figure 5(c).

images per strip are used. We attribute to the aforementioned misalignment between local geometries the fact that the MAE of L1B DSMs with respect to lidar tends to be slightly larger and more irregular.

Lastly, Table 1 shows that the MAE with respect to the lidar is higher for Antibes. This last observation is mainly explained by the presence of vegetation and the edges of the buildings, which are not as sharp in the photogrammetric DSM. Both vegetation and buildings are absent in the Morenci landscape.

5. CONCLUSION

We have presented a method to generate large scale images from fragmented push-frame satellite acquisitions. A perfect sensor localization model is generated for the output images. We denote the resulting product L1B$^+$. The method is validated using two SkySat stereo acquisitions of L1B scenes.

The use of L1B$^+$ offers several advantages over SkySat L1B scenes. In this paper we focused on the benefits for 3D reconstruction, which becomes significantly faster and simpler. We avoid the need to handle multiple stereo pairs and to merge a collection of local models, a common drawback in 3D reconstruction from push-frame imagery. The L1B$^+$ products make it possible to reconstruct areas of interest of several km^2 with a single execution of a satellite stereo reconstruction pipeline. We also notice accuracy improvements in the L1B$^+$ derived DSMs, mainly due to the disappearance of any biases caused by the fusion of local models that is necessary with the L1B scenes.

Future work will focus on extending this methodology to assemble images from all three SkySat sensors at once, combining scenes from three multi-image strips instead of one. The generalization capability of the method should also be investigated using push-frame acquisitions from other satellites.

6. ACKNOWLEDGEMENTS

This work was supported by a grant from Région Île-de-France. It was also partly financed by Office of Naval research grant N00014-17-1-2552, MENRT, and Kayrros. This work was performed using HPC resources from GENCI–IDRIS (grants 2021-AD011012453 and 2022-AD011011801R1). We thank Planet for providing the L1B SkySat images.

REFERENCES

Aati, S., Avouac, J.-P., 2020. Optimization of optical image geometric modeling, application to topography extraction and topographic change measurements using PlanetScope and SkySat imagery. *Remote Sensing*, 12(20), 3418.

Akiki, R., Marí, R., de Franchis, C., Morel, J.-M., Facciolo, G., 2021. Robust rational polynomial camera modelling for SAR and pushbroom imaging. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 7908–7911.

Anger, J., Ehret, T., de Franchis, C., Facciolo, G., 2020. Fast and accurate multi-Frame super-resolution of satellite images. *ISPRS Annals*, 5(1), 57–64.

Baillarin, S., Panem, C., Lebegue, L., Bignalet-Cazalet, F., 2010. Pléiades-HR imaging system: Ground processing and products performances, few months before launch. *ISPRS Archives*, 38(7B), 51-55.

Baker, S., Matthews, I., 2001. Equivalence and efficiency of image alignment algorithms. *CVPR*, 1, 1090–1097.

Beyer, R. A., Alexandrov, O., McMichael, S., 2018. The Ames Stereo Pipeline: NASA's open source software for deriving and processing terrain data. *Earth and Space Science*, 5(9), 537–548.

Bhushan, S., Shean, D., Alexandrov, O., Henderson, S., 2021. Automated digital elevation model (DEM) generation from very-high-resolution Planet SkySat triplet stereo and video imagery. *ISPRS Journal*, 173, 151–165.

Briand, T., Facciolo, G., Sánchez, J., 2018. Improvements of the inverse compositional algorithm for parametric motion estimation. *Image Processing On Line*, 8, 435–464.

Cannistra, A. F., Shean, D. E., Cristea, N. C., 2021. High-resolution CubeSat imagery and machine learning for detailed snow-covered area. *Remote Sensing of Environment*, 258, 112399.

d'Angelo, P., Reinartz, P., 2021. Digital Elevation Models from Stereo, Video and Multi-View Imagery captured by small Satellites. *ISPRS Archives*, 43(B2), 77–82.

de Franchis, C., Meinhardt-Llopis, E., Michel, J., Morel, J.-M., Facciolo, G., 2014a. An automatic and modular stereo pipeline for pushbroom images. *ISPRS Annals*, 2(3), 49–56.

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

d'Autume, M., Perry, A., Morel, J.-M., Meinhardt-Llopis, E., Facciolo, G., 2020. Stockpile monitoring using linear shape-from-shading on PlanetScope imagery. *ISPRS Annals*, 5(2), 427–434.

Facciolo, G., de Franchis, C., Meinhardt-Llopis, E., 2017. Automatic 3D reconstruction from multi-date satellite images. *CVPR Workshops*, 57–66.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L. et al., 2007. The shuttle radar topography mission. *Reviews of geophysics*, 45(2).

Gómez, A., Randall, G., Facciolo, G., von Gioi Grompone, R., 2022. An experimental comparison of multi-view stereo approaches on satellite images. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 844–853.

Gong, K., Fritsch, D., 2018. Point cloud and digital surface model generation from high resolution multiple view stereo satellite imagery. *ISPRS Archives*, 42(2), 363–370.

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

Hartley, R., Zisserman, A., 2004. *Multiple view geometry in computer vision*. Cambridge University Press.

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

Marí, R., de Franchis, C., Meinhardt-Llopis, E., Anger, J., Facciolo, G., 2021a. A generic bundle adjustment methodology for indirect RPC model refinement of satellite imagery. *Image Processing On Line*, 11, 344–373.

Marí, R., de Franchis, C., Meinhardt-Llopis, E., Facciolo, G., 2019. To bundle adjust or not: A comparison of relative geolocation correction strategies for satellite multi-view stereo. *ICCV Workshops*, 2188–2196.

Marí, R., de Franchis, C., Meinhardt-Llopis, E., Facciolo, G., 2021b. Automatic stockpile volume monitoring using multi-view stereo from Skysat imagery. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 4384–4387.

Nguyen, N. L., Anger, J., Davy, A., Arias, P., Facciolo, G., 2021. Self-supervised multi-image super-resolution for push-frame satellite images. *CVPR Workshops*, 1121–1131.

Planet, 2021. Planet imagery product specifications.

Qin, R., 2017. Automated 3D recovery from very high resolution multi-view images. *ASPRS Annual Conference*, 12–16.

Sandau, R., Brieß, K., D'Errico, M., 2010. Small satellites for global coverage: Potential and limits. *ISPRS Journal*, 65(6), 492–504.

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

Stucker, C., Schindler, K., 2020. ResDepth: Learned residual stereo reconstruction. *CVPR Workshops*, 184–185.

Triggs, B., McLauchlan, P. F., Hartley, R. I., Fitzgibbon, A. W., 1999. Bundle adjustment—a modern synthesis. *International Workshop on Vision Algorithms*, Springer, 298–372.

Xue, Y., Li, Y., Guang, J., Zhang, X., Guo, J., 2008. Small satellite remote sensing and applications–history, current and future. *International Journal of Remote Sensing*, 29(15), 4339–4372.