Improving the Pair Selection and the Model Fusion Steps of Satellite Multi-View Stereo Pipelines

Alvaro Gómez, Gregory Randall
Facultad de Ingeniería
Universidad de la República, Uruguay
agomez@fing.edu.uy

Gabriele Facciolo, Rafael Grompone von Gioi
Centre Borelli
ENS Paris-Saclay, France

Abstract

Multi-view stereo reconstruction of scenes from satellite images is traditionally performed with a pair-wise stereo-vision approach: (1) multiple views are grouped into pairs, (2) each pair is processed by two-view stereo methods producing an elevation model or point cloud, lastly (3) the pair-wise reconstructions are integrated and filtered to obtain a final result. These steps are organized in a pipeline and the end-to-end performance of reconstructions depends on the behavior of these steps. This work introduces two changes that increase the performance of the reconstructions: a new pair selection approach and a new integration method are presented. The new pair selection replaces commonly used heuristics with a principled criterion that predicts the completeness of a pair based on offline simulations. The presented integration method is based on an iterated bilateral filter. Experiments show that these changes yield a systematic improvement on the performance of the pipeline.

1. Introduction

Multi-View Stereo (MVS) vision aims at reconstructing a 3D scene from multiple 2D views. In satellite imaging, it has been traditionally performed by a pair-wise MVS approach: the views are grouped into pairs and each pair is processed by two-view stereo matching methods, producing an elevation model or point cloud; then all the pair-wise reconstructions are aggregated to obtain a final result [7, 19, 16, 23]. True MVS methods (which reconstruct the scene directly from the whole set of images) are popular for close range imaging [26, 17] but are still seldom used for satellite images as they have not shown significantly better results [29, 13] or are too computationally expensive to be applied to large scale images [9, 18].

In pair-wise MVS, given a set of N images taken from a scene, $N(N-1)$ ordered pairs can be considered for stereo reconstruction. For each pair, a Digital Surface Model (DSM) of the scene is determined. The final MVS reconstruction of the scene can then be obtained by the integration of all the computed DSMs. The quality of the final reconstruction is determined by the quality of the pair-wise DSMs, which depend on several factors such as the orientations of the views of a pair and changes in the acquisition conditions between the images, among others. Besides the stereo matching step, two other steps are crucial in a pair-wise MVS pipeline to achieve a good reconstruction: (a) the selection of the best pairs to run the pair-wise pipeline and (b) the final integration of the resulting DSMs.

Regarding the pair selection step, multiple factors may influence the quality of a pair and it is hard to identify all of them and tell their relative importance. This difficult task has been traditionally tackled by designing heuristics that take into account the metadata of the images [11, 6]. In [25] a supervised machine learning approach was proposed to derive a quality indicator from the metadata of a pair.

On the other end of the pair-wise MVS pipeline, different methods can be applied for the integration (also called fusion or aggregation) of the information of the computed DSMs. Averaging is the most basic approach; but integration by the median is usually preferred as it takes into account the presence of outliers in the DSMs, as can be seen in a recent review on the matter [21].

This article focuses on the pair selection and the DSM integration steps and presents two contributions to enhance the performance of a satellite MVS pipeline. Firstly, for the pair selection, an approach based on the simulation of image and camera model pairs is presented. Synthetic stereo pairs are simulated under all possible geometric configurations on the hemisphere surrounding an artificial scene and the stereo reconstruction quality can be assessed for each pair. This pre-computed quality is then used as a proxy for the quality of real pairs of images. Secondly, for the integration step, an approach based on the bilateral filtering [27] is presented. Contrary to the most commonly used per-pixel median, the approach allows to better integrate DSMs considering the spatial coherence of different properties of the

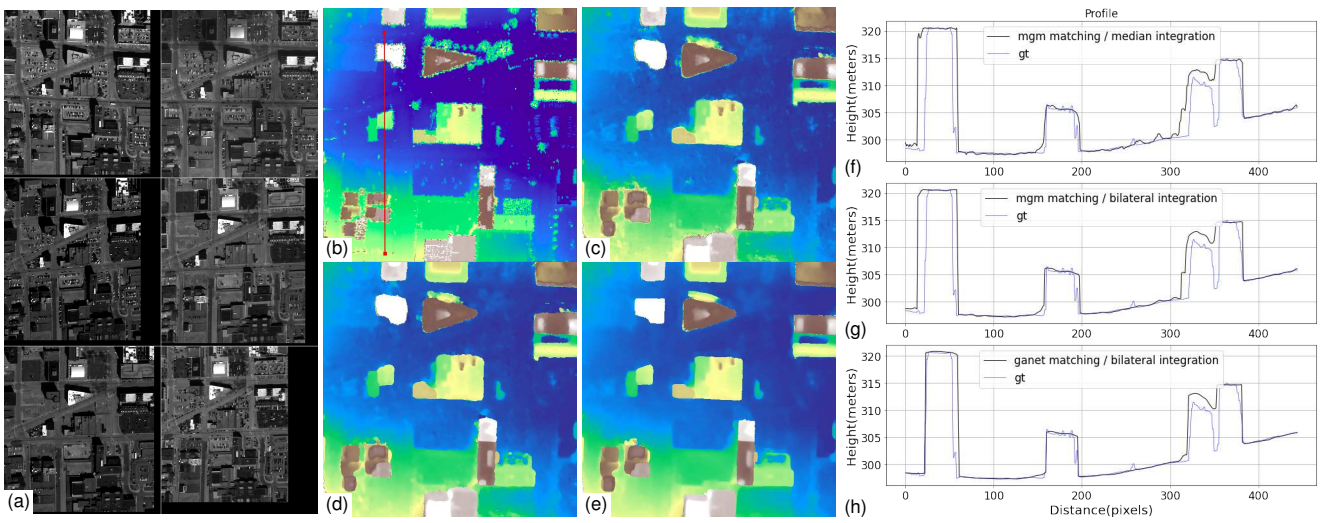

Figure 1. (a) Six images of a region of the Omaha dataset. (b) Ground truth (GT) height map of the region. MVS reconstructions using all 30 stereo pairs: (c) DSMs computed using the MGM [10] correlator and integrated by the median of the DSMs, (d) DSMs computed using the GANet [28, 13] correlator and integrated by the median of the DSMs, (e) DSMs computed using the GANet correlator and integrated by bilateral filtering. Profiles corresponding to the red line: (e) from images b and c, (f) from images b and d, (f) from images b and e.

data such as height, gray level, etc. The method produces a spatial regularization effect, without affecting the borders of the structures as can be seen in Figure 1.

This paper is organized as follows: Section 2 introduces the MVS pipeline used in this work. Section 3 discusses the pair selection alternatives and introduces a new criterion based on simulation. Section 4 presents the DSM integration step. Experimental results are presented in Section 5 and Section 6 concludes and holds the final remarks.

2. The satellite MVS stereo pipeline

The S2P[1] pipeline [7] was used for the experiments in this work. The pipeline input is a stereo pair of images with their respective Rational Polynomial Coefficients (RPC) camera models, which are simplified models for the pushbroom cameras used in satellites [12].

Input images are cut into small tiles. Tiling allows to locally approximate the pushbroom sensor by an affine camera model with a small error, which enables the use of well established stereo rectification and matching methods [7]. Image correspondences are computed on the rectified images with a stereo matching algorithm. The computed correspondences are then triangulated to produce a georeferenced 3D point cloud and a Digital Surface Model (DSM). Lastly, the results for all the tiles are combined to produce the whole image DSM. When multiple images are available, the pipeline can be applied on multiple pairs and the resulting DSMs are integrated to obtain a final DSM.

The experiments in this work are conducted on S2P, but the proposed changes are generic and can be applied to other pipelines as well. We use by default MGM [10] but other correlators are available such as SGM [14], or can be adapted, such as GANet [28] as shown in [13].

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

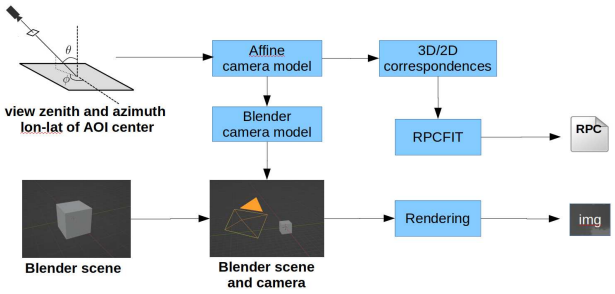

Figure 2. Block diagram of the simulator. Please refer to the text for the description of the blocks and the flow of data.

3. Pair selection for multi-view stereo

In pair-wise MVS, it is well known that the DSM aggregation improves in general the completeness [22, 11]. A new stereo pair may give information of an occluded part of the scene. However, if a DSM computed from a bad pair is included, the result may degrade. This issue along with the fact that the number of possible pairs grows as $O(N^2)$, with N the number of images, makes necessary to pre-select the best pairs to be used for the reconstruction.

In [11] a simple heuristic based on the images metadata was proposed: images in the pair must have an incidence angle smaller than 40º, the angle between views should be in the range [5º, 45º], preferably around 20º and pairs with near acquisition dates are preferred. In [6] pairs with angle

between views in the range [15º, 25º] are preferred for urban and industrial areas.

Here we present a method to empirically map the relation between the orientation of the views and the reconstruction quality through simulation. The simulation tool produce views of a 3D scene from multiple orientations generating images along with RPC models suitable for a pair-wise stereo pipeline. The stereo reconstructions can be assessed by comparing to the known altitude of the scene. This enables to pre-compute a map that encodes the reconstruction quality in relation to the incidence angles of the views with the vertical (or zenith angles) and the intersection angle (angle between views) of any pair of views sampled from the hemisphere surrounding the scene. This map acts as a proxy for the quality of real pairs and can be used to sort the pairs in a more funded way than the previous heuristics.

3.1. Image and RPC simulation tool

Starting from a pre-built 3D scene, the longitude and latitude coordinates of the scene center and the orientation of the view, the simulation tool generates an image and a corresponding RPC camera model suitable to be used in a satellite stereo pipeline. The simulator uses Blender [5] as the 3D engine to render the views. Blender is launched and configured automatically through Python scripts.

Figure 2 presents a block diagram of the simulation tool. Given a scene and a view direction, an affine camera model is determined. The affine camera model is a sensible approximation of a real satellite projection for a small area of interest (AOI) [8]. This model gives corresponding 3D/2D coordinates between the volume of interest (AOI plus height range) and the image. The correspondences are then used to adjust an RPC camera model using the RPCFIT tool [2], which fits an RPC model to the 3D/2D correspondences through a regularized least squares minimization.

In order to render the image of the scene, a camera model, compatible with the affine camera model, is created in Blender. Figure 3 shows examples of the simulation tool with two different scenes: a simple one with a cylinder on a flat surface, and an artificial urban scene.[2] The simulation tool is available at `https://github.com/zemogoravla/simsatool`.

3.2. Stereo reconstruction from simulated image-RPC pairs

The simulator tool allows to draw any pair of views in the hemisphere surrounding a 3D scene. With the generated image pair and their corresponding RPC it is possible to compute a stereo reconstruction with a satellite pipeline and evaluate the reconstruction against the ground truth (GT) altitude of the scene. This enables to empirically study and map the relation between the orientation of the views and the quality of the 3D reconstruction of a pair.

[2]Urban scene downloaded from https://open3dmodel.com/. Accessed on October 2022.

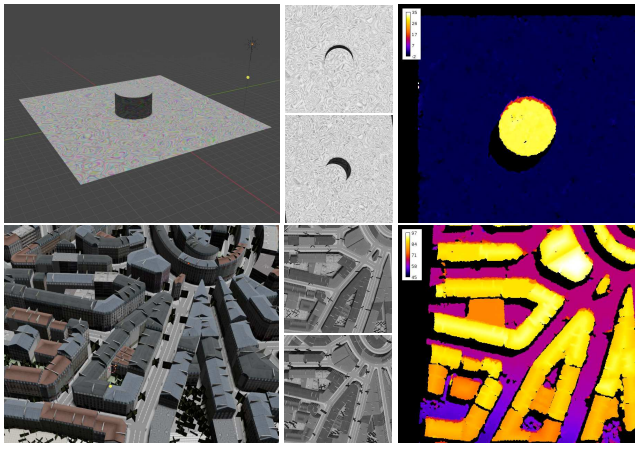

Figure 3. Results of the simulation tool with two different 3D scenes. From left to right: a view of the 3D scene, a stereo pair generated with the simulation tool, and the DSM reconstructed with the S2P pipeline. Above: Cylinder scene. The scene is composed of a cylinder with a radius of 25m and a height of 30m. Surfaces have a random texture. Below: Artificial city scene.

We sampled the hemisphere over a 3D scene and generated pairs of image-RPC from those positions. Figure 4 shows the distribution of the considered reference and secondary views in the hemisphere over the scene. In the plots, the sampled reference views are depicted with a square dot and secondary views with circle dots. Relative orientations are considered with the reference view in zero azimuth. A reference-secondary pair keeps the same relative orientation when a vertical rotation is applied, so should give similar results. Texture or noise can favor some orientations over others. To smooth out these effects, six orientations are considered for each reference-secondary pair. The results are computed as the median over the six cases.

The DSMs computed from the sampled pairs were compared against the GT DSM to assess the reconstruction performance for the stereo pairs. The completeness (COMP) metric, defined as the proportion of the evaluated pixels where the altitude of the computed map differs from the GT less or equal than $z\_tol = 1m$, was considered for the tests.

The completeness is a comprehensive metric traditionally used for the evaluation of satellite stereo reconstructions [4, 3]. Among the evaluated pixels (i.e. with GT information) there are two types of errors in a reconstructed DSM: (a) *Invalid* pixels where the altitude could not be computed, (b) *Bad* pixels where the computed altitude differs from the GT more than a given threshold. *Invalid* pixels are places with incoherent disparities between left and right disparity maps on the stereo matching step. These are mostly caused by occlusions. *Bad* pixels may arise due to matching or triangulation errors. In the first case, repeti-

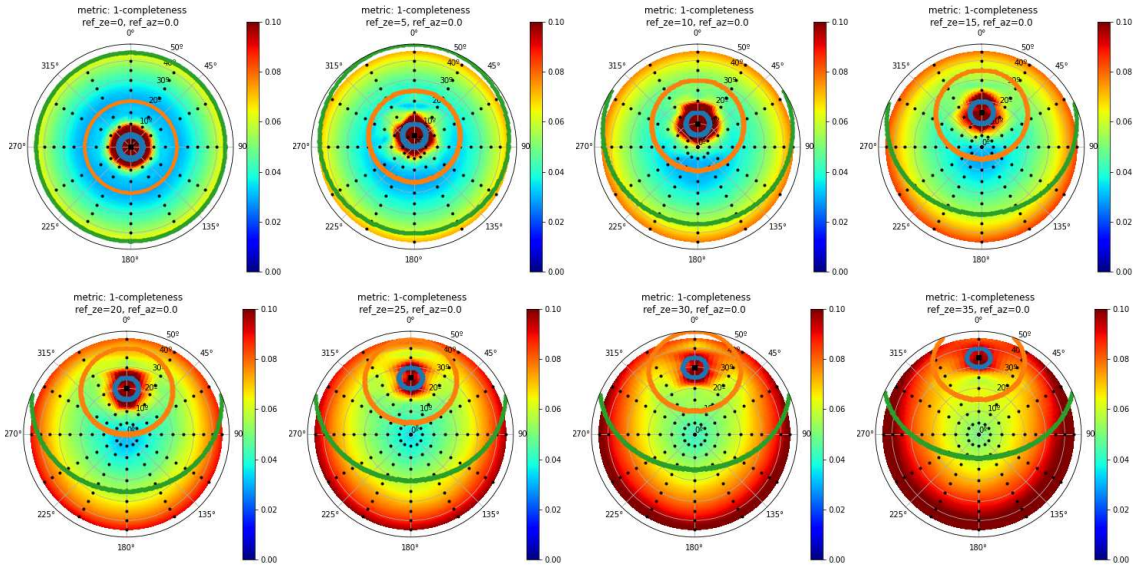

Figure 4. Reconstruction errors of simulations for different reference-secondary image orientations. The square represents the reference view and circular points represent the tested secondary views. Metric 1-COMP is shown for increasing zenith angle of the reference view. Blue corresponds to small errors while red indicates large errors. The blue, orange and green curves indicate the positions in the hemisphere for views 5º, 20º and 45º apart from the reference, respectively.

tive textures may cause coherent matching at a wrong position. Regarding triangulation, the angle between views is the main factor that determines the uncertainty of this step. Small angles between views result in a worse conditioning of the triangulation, which amplifies the small matching errors. On the other hand, an off nadir view implies a foreshortening in one direction causing an anisotropic loss of resolution in the image. Indeed, as shown in [20], the relative position of the views in the hemisphere surrounding the scene entails an affine deformation between the images that may bedevil the matching. *Totalbad* is defined as the sum of *Bad* and *Invalid* pixels and is the complement of the completeness *Totalbad = 1 - COMP*.

Stereo reconstruction is affected in a complex manner by all these factors. The analytical study of all these contributions and interactions is hard, thus simulation becomes a good alternative to tackle this problem.

Figure 4 illustrates the reconstruction error as a function of the reference-secondary relative image orientations. These results are computed using the cylinder scene, but similar results are obtained with more complex scenes. In each case the reference image—the black square—is positioned in a certain zenith angle and the secondary image—the black circular dots—are positioned in a sampling of the hemisphere. Intermediate values are obtained by interpolation of the calculated values at the sampled positions. Blue corresponds to small errors while red indicates large errors.

3.3. MVS pair selection based on simulation results

Given a set of N real satellite images taken from the same region, there are $N \times (N-1)$ possible ordered pairs. For each candidate pair, we can estimate the reconstruction error (as $1 - \text{COMP}$) by querying the orientations of the real images in the pre-computed error maps of Figure 4. This provides an ordering for the integration of the DSMs reconstructed from the pairs. This ordering based on the completeness obtained from the simulation acts as a proxy for the true completeness, which cannot be computed in a real scene where the GT is not available.

4. DSM integration

Multiple strategies to integrate DSMs have been proposed. A recent review [21] presents an extensive list of methods. The most common way of integrating a set of DSMs is to apply a per-pixel median of the heights in the set of DSMs. This usually yields a robust estimation and removes most outliers in the DSMs. However, this pixel-wise approach does not introduce spatial coherence.

In this work an approach based on the bilateral filter [27] is used. The method is related to the one presented in [24] in the sense that both try to include a spatial regularization inspired on the bilateral filter. In [24] for each pixel an irregular region around it is determined considering spatial and color proximity and then a median of the values of the DSMs is applied on that region. Instead, we directly apply a bilateral filter to the samples in the DSMs. The bilateral filter framework allows to robustly integrate the spatial information along with other available sources of information that can regularize the final integrated DSM. Typically, the framework can integrate not only the height of the DSMs and the gray level or color of a reference image, but

also other features as a semantic segmentation or confidence maps if available.

The bilateral filter framework is applied in an iterative scheme. This allows to gradually refine the solution. Using progressively more restrictive ranges for the height allows to focus on the height samples that are close to the previous estimation and are then probably more accurate. The bilateral filter integration of a set of L DSMs for pixel i at iteration n is computed as

$$B[i] = \frac{1}{\nu(i)} \sum_k \sum_{j \in [w_{s_n}, w_{s_n}]} W[k][i,j] L[k][i-j], \quad (1)$$

where

$$W[k][i,j] = e^{-\frac{|j|^2}{2.{s_n}^2}} e^{-\frac{|L[k][i-j]-D[i]|^2}{2.{r_n}^2}} e^{-\frac{|I[i-j]-I[i]|^2}{2.{c_n}^2}}. \quad (2)$$

Here k is an index on the DSMs list, j is an index on the spatial neighbors of pixel i, r_n is the sigma of the Gaussian that determines the height range neighborhood, s_n is the sigma of the Gaussian that determines the spatial neighborhood and c_n is the sigma of the Gaussian for the gray/color value neighborhood on iteration n in all cases, and the normalization factor is

$$\nu(i) = \sum_k \sum_{j \in [w_{s_n}, w_{s_n}]} W[k][i,j]. \quad (3)$$

Algorithm 1 shows the main steps of the method. The inputs are a list of registered DSMs L, a reference gray/color level image I and lists of sigmas to be applied in each iteration to weight the contribution of neighbors to the integrated altitude of each pixel (S: proximity, R: altitude similarity, and C: gray/color value similarity). The image I can be one of the images from the stereo pairs used to compute the DSMs (ortho-rectified to match the DSMs). The integration D is initialized by the per-pixel median of the set of DSMs. For each iteration, (a) the DSMs in L are registered in height to D (shift of each DSM to match the median height of D), (b) the integrated altitude of each pixel is computed.

Algorithm 1: Iterative bilateral DSM integration

input : List of DSMs: L = [DSM[k], k:0...K-1]
Ortho-rectified Reference image: I
Number of iterations: N
List of range sigmas: R=[r_n, $n : 0...N-1$]
List of spatial sigmas: S=[s_n, $n : 0...N-1$]
List of color sigmas: C=[c_n, $n : 0...N-1$]

output: Integrated DSM: D

$D \leftarrow pixel\_wise\_median(L)$
for n *in* $0...N-1$ **do**
  $L \leftarrow [register\_in\_height(L[k], D), for\ k\ in\ 0...K-1]$
  for *each pixel i* **do**
    $B[i] \leftarrow$ as in equation (1)
  $D \leftarrow B$

5. Experiments

Experiments were conducted to test the behavior of the S2P pipeline when changing the pair selection step as proposed in Section 3 and the integration step as proposed in Section 4. We used three datasets, consisting on satellite images from the Multiple View Stereo Benchmark for Satellite Imagery (MVS3D) [4] and the US3D dataset [3].

The MVS3D is a set of 47 satellite images of Buenos Aires (MVS), Argentina. The corresponding GT DSMs are derived from an airborne Lidar acquisition (from a different date than the satellite images) of the same region. The US3D dataset consists of 26 WorldView-3 target-mode panchromatic images collected between 2014 and 2016 over Jacksonville (JAX), Florida and 43 WorldView-3 target-mode panchromatic images collected between 2014 and 2015 over Omaha (OMA), Nebraska. Semantic labels and an airborne Lidar are also available. The Lidar, acquired at a different date than the satellite images by the USGS, is used to derive the GT DSMs.

For our evaluation, 5 subregions from each of the datasets are considered. In each subregion, a set of 6 images is considered in order to allow a tractable pairwise analysis. This gives a set of 30 ordered pairs for each subregion. Images in each set span a small time interval (same day or some days apart) to avoid seasonal changes that could hinder the study.

We tested with two different matching algorithms in the S2P pipeline (MGM[10] and GANet [28]) to show that the presented improvements are rather independent of the used method. In order to evaluate the performance of the different approaches two metrics were considered [4, 3]: (a) Completeness (COMP): Proportion of evaluated pixels where the altitude of the computed map differs from the GT less or equal than $z\_tol = 1m$. (b) Accuracy as the Median Absolute Error (MAE) between computed and GT maps considering only the pixels that have valid information in both maps.

5.1. Analysis of the pair selection strategy.

To analyze the usefulness of the simulation for pair selection, we study if the simulation proxy ranks the pairs in a better way than the commonly used heuristics [11]. This is done by comparing the reconstructed DSMs against the GT.

Table 1 presents the correlation results for the pair rankings given by the heuristic (described in Section 3) and the presented simulation proxy, compared to the rankings obtained by evaluating the true reconstructions. The analy-

	Bad			Invalid			Totalbad = 1 - COMP		
Region	**20º heuristic**	**Cylinder**	**City**	**20º heuristic**	**Cylinder**	**City**	**20º heuristic**	**Cylinder**	**City**
OMA_203	-0.19 (0.91)	**0.68 (<0.01)**	-0.08 (0.71)	**0.25 (0.03)**	**0.69 (<0.01)**	**0.63 (<0.01)**	0.15 (0.13)	**0.30 (0.01)**	0.09 (0.26)
OMA_247	-0.36 (1.00)	**0.78 (<0.01)**	-0.13 (0.82)	-0.05 (0.64)	**0.30 (0.01)**	**0.26 (0.03)**	-0.20 (0.93)	-0.01 (0.53)	-0.28 (0.98)
OMA_251	-0.41 (1.00)	**0.72 (<0.01)**	-0.13 (0.81)	0.12 (0.21)	**0.50 (<0.01)**	**0.43 (<0.01)**	-0.05 (0.63)	0.19 (0.10)	-0.10 (0.75)
OMA_287	-0.40 (1.00)	**0.68 (<0.01)**	-0.08 (0.68)	0.00 (0.50)	**0.47 (<0.01)**	**0.46 (<0.01)**	-0.21 (0.91)	0.00 (0.50)	-0.03 (0.56)
OMA_353	-0.44 (1.00)	**0.74 (<0.01)**	-0.14 (0.86)	**0.29 (0.01)**	**0.66 (<0.01)**	**0.65 (<0.01)**	-0.09 (0.74)	0.12 (0.19)	-0.10 (0.78)
JAX_156	-0.07 (0.71)	**0.68 (<0.01)**	0.17 (0.10)	0.08 (0.28)	**0.24 (0.03)**	**0.38 (<0.01)**	0.07 (0.31)	**0.38 (<0.01)**	-0.01 (0.54)
JAX_165	-0.08 (0.73)	**0.72 (<0.01)**	0.20 (0.08)	0.08 (0.29)	**0.31 (0.01)**	**0.37 (<0.01)**	**0.28 (0.02)**	**0.35 (<0.01)**	**0.39 (<0.01)**
JAX_214	-0.10 (0.76)	**0.58 (<0.01)**	0.14 (0.17)	0.14 (0.17)	**0.26 (0.03)**	**0.41 (<0.01)**	0.19 (0.09)	**0.28 (0.02)**	**0.30 (0.01)**
JAX_251	-0.06 (0.68)	**0.70 (<0.01)**	0.16 (0.11)	0.05 (0.35)	**0.32 (0.01)**	**0.44 (<0.01)**	**0.28 (0.02)**	**0.43 (<0.01)**	**0.31 (0.01)**
JAX_264	-0.10 (0.77)	**0.58 (<0.01)**	0.07 (0.29)	0.07 (0.31)	**0.28 (0.02)**	**0.38 (<0.01)**	0.17 (0.10)	**0.35 (<0.01)**	0.02 (0.43)
MVS_001	**0.28 (0.02)**	**0.80 (<0.01)**	**0.60 (<0.01)**	-0.22 (0.95)	**0.44 (<0.01)**	**0.65 (<0.01)**	**0.45 (<0.01)**	**0.27 (0.02)**	**0.37 (<0.01)**
MVS_002	**0.36 (<0.01)**	**0.86 (<0.01)**	**0.63 (<0.01)**	-0.11 (0.80)	**0.43 (<0.01)**	**0.64 (<0.01)**	**0.65 (<0.01)**	**0.79 (<0.01)**	-0.11 (0.80)
MVS_003	**0.30 (0.01)**	**0.88 (<0.01)**	**0.58 (<0.01)**	-0.25 (0.97)	**0.41 (<0.01)**	**0.64 (<0.01)**	**0.43 (<0.01)**	0.15 (0.13)	**0.53 (<0.01)**
MVS_004	**0.28 (0.02)**	**0.86 (<0.01)**	**0.57 (<0.01)**	-0.21 (0.94)	**0.40 (<0.01)**	**0.62 (<0.01)**	**0.54 (<0.01)**	**0.30 (0.01)**	**0.48 (<0.01)**
MVS_005	0.17 (0.11)	**0.75 (<0.01)**	**0.48 (<0.01)**	-0.26 (0.97)	**0.37 (<0.01)**	**0.59 (<0.01)**	**0.45 (<0.01)**	**0.38 (<0.01)**	0.00 (0.49)

Table 1. Analysis of the pair rankings given by the heuristic and the simulation compared to the rankings given by the true reconstructions. For a given metric each cell shows the Kendall-tau correlation and its p-value. For example, in the cell corresponding to Bad/Cylider/OMA_203, the ranking by metric Bad of the pairs from that region (metric computed comparing the DMSs against the GT) and the ranking by metric Bad for the cylinder based simulation, have a Kendall-tau correlation of 0.68 with p-value < 0.01. The water and vegetation pixels were masked out for this analysis. Cells highlighted in bold correspond to correlations with p-value < 0.05

sis is repeated for each of the error metrics (Bad, Invalid and Totalbad). For a given metric each cell in the table shows the Kendall-tau (KT) rank correlation coefficient and its corresponding p-value [15, 1]. Simulations are made on both scenes shown in Figure 3 (i.e. Cylinder and City).

As mentioned in Section 3, the errors of a stereo reconstruction are comprised of *Bad* and *Invalid* pixels. While the *Bad* pixels are more related to the orientation of the views, *Invalid* pixels have a strong relation to the geometry of the scene (e.g. occlusions are related to the contents and the spatial relation of the objects in the scene). Results on Table 1 show that for the simulation with the Cylinder scene, there is a significant correlation between the rankings for both the Bad and Invalid metrics on simulations using the Cylinder and City scenes. This simple scene correctly captures the main error components given the view orientations and can rank the pairs in a similar way as with the real images. In particular, the simulation on the Cylinder scene presents a strong correlation for the Bad metric, which is mostly related to the views and not to the scene. Bad pixels are mostly related to the view orientations and not to errors in the matching step and the simple Cylinder scene allows to observe these errors independently of problems that a more complex scene could introduce. In the case of the Invalid metric, the Cylinder has still a positive and significant correlation but the City scene, with a more complex structure, represents better the inter-occlusions of a urban scene.

From Table 1 we see that the simulation based pair selection rightly predicts the best ordering in relation with the *Bad* and the *Invalid* number of pixels, but is less conclusive for the number of *Totalbad* pixels. We shall note that Bad and Invalid are antagonistic metrics. A large angle between views reduces the uncertainty for the triangulation while causes large occluded regions. This antagonistic relation and the dependence on the scene layout explains that the correlation for *Totalbad* with the ideal ordering is not as strong as the correlation of its components. We posit that simulating using an adequate layout adapted to each particular scene (a priori unknown), would improve this correlation. Despite this limitation, the results show that the pair selection method using a very simple simulation model (Cylinder) gives good results in mean, as illustrated in Figure 6, and works better than the currently used heuristic method.

The experiments show that it is possible to develop a better (founded) pair selection strategy than the currently used heuristic [11]. The simulation tool allows to consider all the possible configurations of incidence and angles between views, related to the two error types mentioned in Section 3. Overall, the correlation of the heuristic strategy with the different metrics is rather disappointing, except for the case of the MVS sets, on which it was fine-tuned [11]. This seems to indicate that the existing heuristic miss some relevant cases. For instance, note that in Figure 4 the heuristic of 20º preference for the angle between views [11] is confirmed by the first plots. But the simulation reveals that as the incidence angle of one of the views grows, it is preferable to have the other view near the nadir even if the angle between views moves away from 20º. Less error is found when the secondary view has a similar azimuth to the reference (that is, moving from the reference to the nadir). A secondary view near the nadir in the same azimuth as the reference does not increase the occlusions and has maximum resolu-

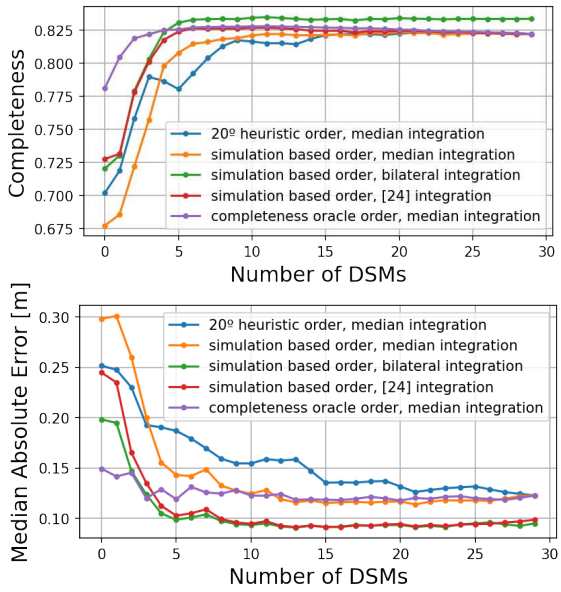

Figure 5. Progressive integration of DSMs for region JAX_156. Completeness (top) and MAE (bottom) evolution when integrating a growing number of DSMs. Purple curve corresponds to an integration by the median and an oracle ordering. Curves in blue, orange and green correspond to the C1, C2 and C3 configurations of the pipeline respectively. Red curve corresponds to an implementation of the method in [24]. Refer to the text for a complete description.

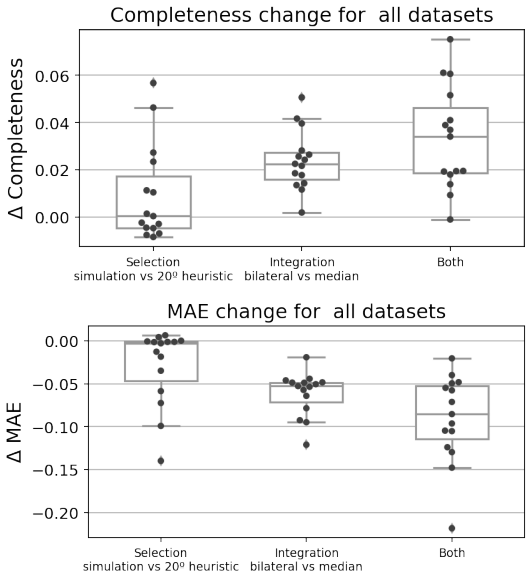

Figure 6. Incremental results between configurations C1, C2 and C3 (see text). For each region in the datasets, the top 5 DSMs are integrated and the resulting DSM compared to the GT. Left: C2 - C1 (pair selection by simulation vs. heuristics), Center: C2 - C1 (integration by bilateral filter vs. integration by the median). Right: C3 - C1 (both improvements vs. the original baseline pipeline). Note how each contribution increases the global performance.

tion as it minimizes the foreshortening. That view will be better than one 20º apart from the reference with different azimuth (e.g a view to a side of the reference with same zenith angle).

The simulation can be used as a pre-computed mapping to estimate the expected completeness and select the pairs in such a way as to minimize the 3D reconstruction error. The analysis of the simulation results shows that a simple scene as the Cylinder can be used to order the pairs.

5.2. Analysis of the end-to-end performance.

In order to assess the end-to-end effects of the presented contributions (pair selection and DSM integration) the following configurations of the pipeline were tested: (C1) Selection by the 20º heuristic, integration by median, as in the current S2P pipeline, (C2) Selection by simulation as proposed in section Section 3, integration by median, (C3) selection by simulation, integration by iterative bilateral filtering as presented in Section 4. Regarding the iterative bilateral filtering, the shown results use a decreasing sigma for the height range of $[2.5, 2.0, 1.5, 1, 0.5]$, spatial sigma of 6 and color sigma of 20% of the gray level range.

Figure 5 illustrates the performance change on one region of the dataset when the contributions of this work are introduced in the pipeline (results for other regions are available in the supplementary material). The graphs show the behavior of the two metrics—completeness and median of the absolute error—for the region as the number of integrated DSMs is increased according to different ordering criteria. Purple curve depicts an integration with the median and an "oracle" ordering based on the actual completeness computed with an available altitude GT. In the oracle ordering, the next DSM is selected to maximize the completeness of the integration up to the moment. This almost optimal ordering illustrates the common situation where the completeness peak is achieved with the DSMs from a few good pairs and the inclusion of more DSMs degrades the aggregated result. These "toxic" DSMs are the result of bad image pairs. The oracle puts these bad DSMs at the end of the ordering. Blue, orange and green curves correspond to the C1, C2 and C3 configurations respectively. In the example of Figure 5 the introduced methods enhance the completeness and the accuracy allowing to achieve better results with fewer DSMs. Considering, for example, the first five selected DSMs, the selection by the simulation is closer to the ideal selection by the oracle; integration by iterated bilateral filtering adds another performance boost that surpasses the peak performance of the integration by the median.

This trend is general for the ensemble of the datasets as depicted in Figure 6, which shows, for all the tested regions (dots in the figure), the variation in the reconstruction met-

			COMP				MAE			
Selection 5/30	**Matching method**	**DSM Integration**	**Jacksonville**	**Omaha**	**Buenos Aires**	**All**	**Jacksonville**	**Omaha**	**Buenos Aires**	**All**
20º Heuristic	MGM	Median	0.700	0.811	0.720	0.744	0.306	0.231	0.285	0.274
By simulation	MGM	Median	0.733	0.811	0.715	0.753	0.225	0.227	0.284	0.245
By simulation	MGM	Bilateral	**0.747**	**0.829**	**0.732**	**0.770**	**0.199**	**0.180**	**0.255**	**0.212**
20º Heuristic	GANet	Median	0.695	0.832	0.723	0.750	0.373	0.232	0.306	0.304
By simulation	GANet	Median	0.725	0.830	0.718	0.758	0.296	0.246	0.307	0.283
By simulation	GANet	Bilateral	**0.736**	**0.838**	**0.731**	**0.769**	**0.275**	**0.228**	**0.287**	**0.263**

Table 2. Results for the whole satellite pipeline on the tested data sets for configurations C1, C2 and C3. The results are the average of the metrics on the datasets. In all cases the comparison is against the GT and using the best five pairs chosen either by the heuristics or by the simulation tool. Note how the integration by bilateral filter improves the completeness (COMP) over the median approach both when using the MGM or the GANet matching methods. The same behavior is observed for the Median of the absolute differences (MAE).

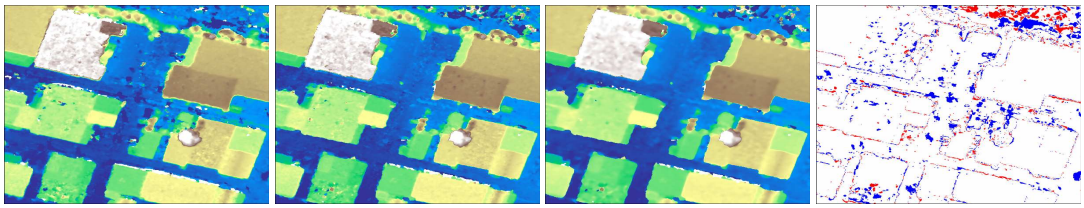

Figure 7. An example of results for a region from the Jacskonville dataset. From right to left the first 3 columns show the reconstruction using the best 5 DSMs and the configurations: (C1) Selection by the 20º heuristic, integration by median, (C2) Selection by simulation, integration by median, (C3) selection by simulation, integration by iterative bilateral filtering. All results use the MGM stereo matcher. Last column graphically compares the completeness difference between (C3) and (C1): blue color are correctly reconstructed pixels (height error $< 1m$) by (C3) and badly reconstructed by (C1); red color are badly reconstructed pixels by (C3) and correctly reconstructed by (C1). Note that the improvements –in blue– prevail and are concentrated on the edges of the structures, with the exception of an upper right region with vegetation.

rics when the presented improvements are introduced into the pipeline. The three configurations defined before (C1, C2 and C3) are considered: C1 is the baseline, C2 changes the selection of pairs with respect to C1, and C3 changes the integration step with respect to C2. Figure 6 shows the error metrics change considering (C2-C1), (C3-C2) and (C3-C1) to evaluate the contribution of each proposed improvements in the global performance. The boxplots show that each contribution produce a consistent improvement in the reconstructions both in completeness and accuracy.

As mentioned in Section 4, the presented bilateral filtering integration method is related to [24]. The method, hereafter called bilateral median (BM), was implemented in order to compare it with our proposal. Figure 5 compares, for a given image, the integration evolution with the BM (red) and with the bilateral filtering (green), using the same parameters. BM exhibits very good results with the first few DSMs but falls behind bilateral filtering integration as the number of DSMs increase. This evolution is similar for all the tested regions (results for other regions are in the supplementary material). While color range and spatial regularization are common to both methods, the ability to take into account the height range with decreasing sigmas is key to integrate the best of all available DSMs. Note that the implementation and parameter values for the BM method might differ from the actual method in [24].

Table 2 summarizes the results obtained for configurations C1, C2, and C3, averaged by site. Performance gain is mainly due to the integration by the bilateral filter. Figure 6 shows that the contribution of the selection is positive in mean, in spite of the fact in Table 2 that for some sites the simulated based selection is not optimal. Both contributions combined improve the overall performance of the pipeline in terms of completeness and accuracy and this observation persists regardless of the matching method used (MGM or GANet). Figure 7 shows that the completeness improvements of the integration method are concentrated on the borders of the structures like buildings. This contributes to a better definition and fidelity to the GT of the reconstructed 3D structures as seen also in Figure 1.

6. Conclusions and future work

In this paper we present two alternative steps for the MVS satellite pipelines: the method to select the pairs to be used and the method to integrate the resulting DSMs. Experiments show that both improve the completeness of reconstructed DSMs and reconstruction accuracy. The results are consistent for two different stereo matching methods. Integration by bilateral filtering systematically attains better results compared to the classic median integration. The pair selection based on the results of a simulation has achieved encouraging results. In the presented approach, the orientation of the views are considered. The tool can be further improved by including the sun position in the simulation.

References

[1] Hervé Abdi. The kendall rank correlation coefficient. *Encyclopedia of measurement and statistics*, 2:508–510, 2007.

[2] Roland Akiki, Roger Marí, Carlo De Franchis, Jean-Michel Morel, and Gabriele Facciolo. Robust rational polynomial camera modelling for sar and pushbroom imaging. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 7908–7911, 2021.

[3] Marc Bosch, Kevin Foster, Gordon Christie, Sean Wang, Gregory D Hager, and Myron Brown. Semantic stereo for incidental satellite images. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1524–1532. IEEE, 2019.

[4] Marc Bosch, Zachary Kurtz, Shea Hagstrom, and Myron Brown. A multiple view stereo benchmark for satellite imagery. In *2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–9. IEEE, 2016.

[5] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018.

[6] Pablo d'Angelo, Cristian Rossi, Christian Minet, Michael Eineder, Michael Flory, and Irmgard Niemeyer. High resolution 3d earth observation data analysis for safeguards activities. In *Symposium on International Safeguards*, pages 1–8, 2014.

[7] Carlo de Franchis, Enric Meinhardt-Llopis, Julien Michel, Jean-Michel Morel, and Gabriele Facciolo. An automatic and modular stereo pipeline for pushbroom images. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3:49–56, 2014.

[8] Carlo de Franchis, Enric Meinhardt-Llopis, Julien Michel, Jean-Michel Morel, and Gabriele Facciolo. On stereo-rectification of pushbroom images. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 5447–5451, Paris, France, Oct. 2014. IEEE.

[9] Dawa Derksen and Dario Izzo. Shadow neural radiance fields for multi-view satellite photogrammetry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1152–1161, 2021.

[10] Gabriele Facciolo, Carlo de Franchis, and Enric 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.

[11] Gabriele Facciolo, Carlo De Franchis, and Enric Meinhardt-Llopis. Automatic 3D Reconstruction from Multi-date Satellite Images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1542–1551, Honolulu, HI, July 2017. IEEE.

[12] Jacek Grodecki. Ikonos stereo feature extraction-rpc approach. In *ASPRS annual conference St. Louis*, 2001.

[13] Alvaro Gómez, Gregory Randall, Gabriele Facciolo, and Rafael Grompone von Gioi. An experimental comparison of multi-view stereo approaches on satellite images. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 707–716, 2022.

[14] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2008.

[15] Maurice George Kendall. Rank correlation methods. 1948.

[16] Thomas Krauß, Pablo d'Angelo, Mathias Schneider, and Veronika Gstaiger. The fully automatic optical processing system catena at DLR. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pages 177–183, 2013.

[17] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[18] Roger Marí, Gabriele Facciolo, and Thibaud Ehret. Sat-nerf: Learning multi-view satellite photogrammetry with transient objects and shadow modeling using rpc cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1311–1321, 2022.

[19] Zachary M Moratto, Michael J Broxton, Ross A Beyer, Mike Lundy, and Kyle Husmann. Ames stereo pipeline, NASA's open source automated stereogrammetry software. *LPI*, (1533):2364, 2010.

[20] Jean-Michel Morel and Guoshen Yu. Asift: A new framework for fully affine invariant image comparison. *SIAM journal on imaging sciences*, 2(2):438–469, 2009.

[21] Chukwuma J Okolie and Julian L Smit. A systematic review and meta-analysis of digital elevation model (dem) fusion: pre-processing, methods and applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, 188:1–29, 2022.

[22] Ozge C Ozcanli, Yi Dong, Joseph L Mundy, Helen Webb, Riad Hammoud, and Victor 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, June 2015.

[23] Rongjun Qin. Rpc stereo processor (rsp)–a software package for digital surface model and orthophoto generation from satellite stereo imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3:77, 2016.

[24] Rongjun Qin. Automated 3d recovery from very high resolution multi-view satellite images. In *Proceedings of the ASPRS Conference (IGTF) 2017, Baltimore, MD, USA, 12–17 March 2017*, pages 12–16, 2017.

[25] Rongjun Qin. A critical analysis of satellite stereo pairs for digital surface model generation and a matching quality prediction model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 154:139–150, 2019.

[26] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016.

[27] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pages 839–846. IEEE, 1998.

[28] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019.

[29] Kai Zhang, Noah Snavely, and Jin Sun. Leveraging vision reconstruction pipelines for satellite imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.