GROUND TRUTH GENERATION AND DISPARITY ESTIMATION FOR OPTICAL SATELLITE IMAGERY

M. Cournet[1,*], E. Sarrazin[1], L. Dumas[2], J. Michel[1], J. Guinet[2], D. Youssefi[1], V. Defonte[2], Q. Fardet[2]

[1] Centre National d'Etudes Spatiales (CNES), 18 avenue E. Belin, Toulouse cedex 9, France
[2] CS, 5 rue Brindejonc des Moulinais, Toulouse Cedex 5, France

Commission II, WG II/2

KEY WORDS: Stereo Ground Truth, Disparity, Stereo-Matching, 3D, Pandora, Optical Satellite Imagery, CO3D

ABSTRACT:

Several 3D reconstruction pipelines are being developed around the world for satellite imagery. Most of them implement their own versions of Semi-Global Matching, as an option for the matching step. However, deep learning based solutions already outperform every SGM derived algorithms on Kitti and Middlebury stereo datasets. But these deep learning based solutions need huge quantities of ground truths for training. This implies that the generation of ground truth stereo datasets, from satellite imagery and lidar, seems to be of great interest for the scientific community. It will aim at reducing the potential transfer learning difficulties, that could arise from a training done on datasets such as Middlebury or Kitti. In this work, we present a new ground truth generation pipeline. It produces stereo-rectified images and ground truth disparity maps, from satellite imagery and lidar. We also assess the rectification and the disparity accuracies of these outputs. We finally train a deep learning network on our preliminary ground truth dataset.

1. INTRODUCTION

Several 3D reconstruction pipelines are being developed around the world for satellite imagery. One can quote ASP for AMES Stereo Pipeline (Shean et al., 2016), Catena (d'Angelo, Kuschk, 2012), MicMac (Rupnik et al., 2018), RSP for RPC Stereo Pipeline (Qin, 2016), or S2P for Satellite Stereo Pipeline (de Franchis et al., 2014). In the frame of the CNES / Airbus CO3D mission (Lebegue et al., 2020), CNES, the French space agency, is also developing its own pipeline, called CARS (Youssefi et al., 2020). This new multi-view stereo pipeline is focused on robustness and scalability, as it will be used for massive DSM production (Melet et al., 2020).

For all those pipelines, two key factors of the 3D restitution accuracy are the image geometry modelling and the image matching, which is also the more time consuming.

Most of these pipelines implement their own versions of Semi-Global Matching, as an option for the matching step (Hirschmuller, 2008). SGM based solutions won the 2016 IARPA Multi-View Stereo 3D Mapping Challenge (Bosch et al., 2017). SGM based solutions also reached the top 3 entries of the multi-view semantic stereo challenge of the 2019 Data Fusion Contest (Le Saux et al., 2019), as noticed by the winners (d'Angelo et al., 2019).

However, deep learning based solutions already outperform every SGM derived algorithms on Kitti and Middlebury stereo datasets (Menze et al., 2018) (Scharstein, Szeliski, 2002). It means that multi-view stereo pipelines users could be interested in testing the most promising deep-learning approaches for the stereo-matching step (just replacing SGM based solutions by these ones). But these deep learning based solutions need huge quantities of ground truths for training. This implies that the generation of ground truth stereo datasets, from satellite imagery and lidar, seems to be of great interest for the scientific community. It will aim at reducing the potential transfer learning difficulties, that could arise from a training done on datasets such as Middlebury or Kitti.

In this publication, we first present a new pipeline to produce a ground truth dataset from optical satellite imagery and lidar raster. The main outputs of our pipeline are stereo-rectified images pairs, and their corresponding disparity maps. The rectification step is based on a strategy recently introduced by the CARS pipeline (Michel et al., 2020). The disparity estimation relies on a new methodology presented in this paper.

In this work, we also evaluate the main outputs of our ground truth pipeline. We focus on the assessment of the rectification error of the stereo-rectified images pairs; and on the accuracy of the disparity maps.

Finally, we present a use case, in which a preliminary groud truth dataset is used to train a deep-learning stereo-matching algorithm. This algorithm is implemented in our stereo matching framework, called Pandora. Pandora will be publicly available as an open-source software; and it will be the stereo-matching tool of the future CO3D mission ground segment.

2. RELATED WORK

Two satellite stereo datasets have been recently published: the Urban Semantic 3D (US3D) dataset (Bosch et al., 2019), and the SatStereo dataset (Patil et al., 2019). To our knowledge, these two datasets are the only ones that include stereo-rectified images pairs and ground truth disparities, generated from satellite imagery and lidar. These two datasets are great initiatives, as they help promoting machine learning for remote sensing stereo reconstruction.

We understand that both are based on the affine sensor assumption for the stereo-rectification step. This implies that the rectification is done on a small tile basis (around 1000×1000-

*Corresponding author

pixel size). Following the methodology described in (de Franchis et al., 2014), the rectifying similarities are computed from an affine fundamental matrix, that is derived from matches. The US3D uses Rational Polynomial Coefficients (RPC) virtual matches, whereas the SatStereo uses SIFT matches to compute these rectifying similarities. That could explain the difference in the observed rectification errors of the two datasets. As mentioned by Bosch et al., some residual y parallax exists in the US3D stereo-rectified image pairs, that is directly linked to the RPC relative pointing accuracy. As mentioned by Patil et al., the SatStereo dataset achieves an average rectification error within half a pixel. Their rectifying similarities computation is based on SIFT matches. That makes this rectification process insensitive to the RPC relative pointing accuracy, but SIFT dependant, and therefore image content dependant. Note that SatStereo is stitching rectified tiles, to produce larger rectified chips (around 5000×5000-pixel size).

Then, in the SatStereo, the disparity estimation is based on a *colocalization* using the lidar. Each point of the left rectified image is mapped into the left original sensor image coordinates using the inverse left rectifying similarity. Then, the original sensor image corresponding point is localized onto the aligned lidar. Next, this 3D point is projected into the right original sensor image coordinates; and into the right rectified image coordinates using the right rectifying similarity. The disparity is finally computed by subtracting the column indices of the positions in the right and left rectified images. (Notations: In our paper, the RPC model projection function gives the 2D image position of a 3D space point; while the localization function is its inverse. The colocalization function gives the 2D image position in image B of a 2D image position in image A, using the localization function of image A and the projection function of image B)

The validation of the disparity accuracy is a tricky task. (Patil et al., 2019) use human annotated tie points, measuring an average disparity error higher than 1.2 pixel. The US3D ground truth disparities are given as integers, and the disparity accuracy is not evaluated. Obtaining a subpixel disparity accuracy is necessary for remote sensing 3D reconstruction, but it seems to be a great challenge.

Our stereo-rectification process differs from the US3D and SatStereo ones, as it does not rely on the affine sensor assumption. Thus easing the ground truth pipeline with no specific tile borders management. It also combines RPC with a SIFT-based correction to improve the rectification accuracy. Then, we use a novel approach to create the disparity maps and show that it achieves a systematic sub-pixel median absolute disparity error on the building class.

3. REMOTE SENSING GROUND TRUTH GENERATION

Here we present a new methodology to produce a ground truth dataset from optical satellite imagery and lidar raster. The inputs are two optical satellite images acquired on the same area with different viewing angles, their RPC models, and a lidar ground truth in raster format. The outputs are stereo-rectified images and ground truth disparity with occlusion mask. The output dataset can then be used to assess stereo-matching algorithms, and/or to train machine learning based solutions.

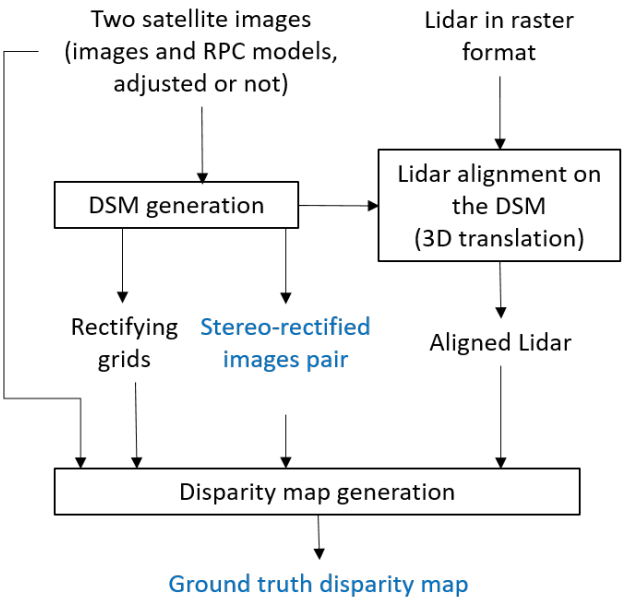

Figure 1. Ground truth pipeline overview on small areas of interest (where bundle adjustment can be safely bypassed). The inputs are a satellite imagery pair and lidar. The outputs are stereo-rectified images pair and a ground truth disparity map

3.1 Challenges

Whatever the methodology, some difficulties arise while generating this kind of ground truth dataset from pushbroom images. Here is a non-exhaustive list:

1. the geometric models errors lead to bad alignment between the lidar and the satellite images, and between the images themselves;
2. the pushbroom acquisition complicates the stereo-rectification process, as the epipolar curves are neither straight lines, nor conjugated;
3. the insufficient spatial resolution of the lidar data implies either height or disparity interpolation, according to the selected disparity map generation methodology;
4. using a lidar that is sampled at a finer step than the stereo-rectified images can lead to disparity aliasing;
5. the date difference between the lidar and the images, or even between the images themselves, can generate landscape changes (e.g.: vegetation or buildings changes).

3.2 Steps of this ground truth pipeline

Following (Patil et al., 2019), our ground truth pipeline is based on: i) Data Alignment - Bundle adjustment, ii) DSM generation from satellite imagery, including rectifying grids and stereo-rectified images pair, iii) Data Alignment - Lidar alignment on the DSM, iiii) Disparity map generation from aligned lidar. Note that algorithms used in each step are different than the ones selected by Patil et al.. Our overall pipeline is also more straightforward, as the tile stitching is useless in our case. A pipeline overview is shown Fig. 1.

3.2.1 Data Alignment - Bundle adjustment The ground truth generation pipeline begins by correcting the pointing error of the RPC models in a bundle adjustment step. This geolocation correction can be done either in a relative way, using tie points; or in an absolute way, using also Ground Control Points. The aim of the relative correction is to increase the stereo-rectification accuracy; while the objective of the absolute correction is also to improve the alignment of the satellite images pair on the lidar raster. This bundle adjustment step is done by an in-house tool. It is optional, and can be safely bypassed on small areas of interest, where the affine camera model assumption is valid. In that case, we just apply a translation to the right stereo-rectified image, and a 3D translation to the lidar (Facciolo et al., 2017).

3.2.2 DSM generation from satellite imagery, including rectifying grids and stereo-rectified images pair Then, the pipeline produces a DSM from the satellite images and their RPC geometric models. The aim of this step is two-fold:

1. to generate the stereo-rectifying grids, and the satellite DSM. These two data are intermediary products. The grids will be used in section 4.2.2; while the DSM will be used in section 4.2.1.

2. to produce the stereo-rectified images, that are direct outputs of this ground truth dataset.

The DSM is generated via CARS, a new multi-view stereo 3D reconstruction pipeline. An exhaustive algorithmic description of CARS can be found in (Michel et al., 2020). Here, we briefly introduce this tool, with a quick focus on the CARS rectification process, as the rectified images are part of the outputs of this ground truth pipeline. CARS is based on two main steps: a preparation step and a DSM computation step.

The CARS preparation part aims at computing the left and right rectifying grids, and the disparity search range. Each rectifying grid is regular in the rectified image geometry, and it gives the original sensor image positions. A displacement along a row in the epipolar grids corresponds to 2D displacements along the two epipolar curves in the original sensor images. These 2D displacements in the original sensor images are computed, jointly and iteratively, using the localization and the projection functions of the two RPC models, and a low resolution DSM (e.g.: SRTM). This stereo-rectification process makes it possible to rectify whole pushbroom images, without the need for tiling that is induced by the affine camera assumption. If the bundle adjustment step is bypassed, the right rectifying grid can be corrected using SIFT matches, following the RPC and SIFT combined approach proposed in (de Franchis et al., 2014). In that case, the SIFT correction amounts to applying a translation (or a more complex transformation) in image space to the right rectified image. It aims at reducing the relative pointing error between the two RPC models, in order to improve the stereo-rectification accuracy.

Then, the CARS DSM computation part begins by stereo-matching the stereo-rectified images; next, it triangulates the homologous points found; finally, it rasterizes the 3D point cloud. Note that the stereo-matching step is done while using Pandora (here in the SGM configuration), cf section 5.2.

3.2.3 Data Alignment – lidar alignment on the DSM This step is optional, and it consists in aligning the lidar on the satellite Digital Surface Model (DSM), produced via CARS. This step is needed if the bundle adjustment step did not include any Ground Control Points. The goal of this alignment is to compensate the absolute pointing error of the RPC models. Note that, neither our current bundle adjustment, nor this lidar 3D translation compensate the potential impact of acquisition vibrations, that were not stored in the geometric model.

As shown in (Facciolo et al., 2017) on small areas of interest, where the affine camera model assumption is valid, the pointing error generates a translation in 3D. This implies that, on these small regions, the bundle adjustment can be skipped, in favour of a 3D translation of the output DSM. Note that (Shean et al., 2016) generally bypass the bundle adjustment step for 3D reconstruction using ASP, on WorldView-1 and WorldView-2 images, noticing that a 3D translation of the entire DSM is almost always sufficient.

The lidar alignment consists in applying a 3D translation to the lidar. The x-y translation is estimated via the Nuth and Kääb methodology (Nuth, Kääb, 2011). The z-translation corresponds to the median height difference between the DSM and the lidar (this median being computed on the coherent low slopes between the DSM and the lidar). Note that the x-y translation is applied by shifting the lidar geotransform origin (i.e.: without lidar resampling).

3.2.4 Disparity map generation from aligned lidar The aim of this step is to produce the left and right disparity maps from aligned lidar (the lidar alignment can be optional, cf section 3.2.3). The inputs are a stereo-rectified images pair, the rectifying grids, the aligned lidar, the original sensor images and their (adjusted) RPC models. The outputs are left and right disparity maps, and occlusion mask.

Even if the *colocalization* strategy proposed by (Patil et al., 2019) in the SatStereo is the more intuitive methodology (cf section 2 for a description), it exists other ways for estimating a disparity map from the above listed inputs. Each one presenting its own advantages and shortcomings. Here we propose and evaluate another method, that we call *localization and height to disp* (cf Fig. 2). It follows three main steps:

1. The left disparity map has the same geometry as the left stereo-rectified image. Each pixel (i, j) of the left stereo-rectified image is mapped into the original sensor image coordinates using the left rectifying grid. Then, the original sensor image corresponding point is localized onto the aligned lidar. Next, the corresponding height is stored into the left disparity map at each position (i, j).

2. We apply pixel-wise bias and ratio to the heights stored into the left disparity map. That gives the disparity map in the left stereo-rectified image geometry. Local ratio and local bias are computed for each pixel of the left stereo-rectified image. The local bias is the local height corresponding to a local disparity equals to 0. The local ratio is the local height delta corresponding to a local disparity step of one pixel. These local ratio and local bias are computed while triangulating the corresponding points in the original sensor images.

$$disparity[i, j] = \frac{height[i, j] – bias[i, j]}{ratio[i, j]} \quad (1)$$

With the CARS rectification methodology, using SRTM as input, the bias can be considered as a SRTM elevation

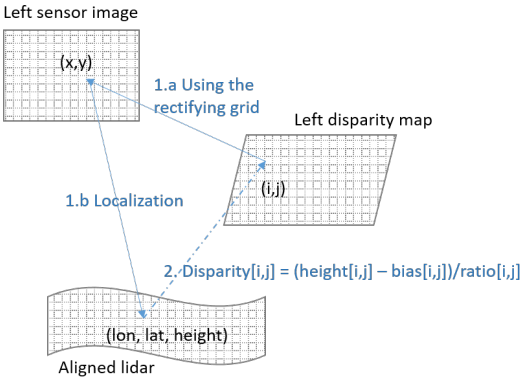

Figure 2. Overview of the disparity map generation process, that we call *localization and height to disp*, cf detailed description in 3.2.4

sampling, whose elevation has been converted to a height above ellipsoid; while the ratio could be approximated by the resolution divided by the stereoscopic angle between the two views. But we remind that we apply bias and ratio computed in a pixel-wise manner. By applying a local ratio, as described above, we make the assumption that the local ratio is valid for every disparity range, which is an approximation.

3. Finally, we perform the same two previous steps on the right stereo-rectified image. That leads to the disparity map in the right stereo-rectified image geometry; and to an occlusion mask thanks to cross-checking.

Note that the step 1 of this *localization and height to disp* process is common with the *colocalization* methodology. They differ in the second step, where the colocalization projects the 3D point into the right image, then in the right rectified image using the right similarity; and finally computes the disparity (cf section 2).

Here is a non-exhaustive list of the shortcomings of these two disparity map generation methods (*colocalization* and *localization and height to disp*). Both depend on the rectification accuracy of the stereo-rectified images. Both depend on the accuracy of the lidar alignment. Both will suffer from the time delta between the lidar and the images. The *colocalization* also directly depends on the relative pointing error of the two (adjusted) RPC models, while our method applies local bias and local ratio that are source of approximations.

Our current pipeline also deals with lidar in raster format. That means that our method depends on the quality of the lidar rasterization. We also interpolate the lidar height during the localization. Note that even if these lidar rasterization and height interpolation could blur the disparity, they could help reducing aliasing, contrary to lidar point cloud sub-sampling. An example of our ground truth outputs and input lidar can be found is Fig. 3. It illustrates few method shortcomings.

4. GROUND TRUTH EVALUATION

In this work, we favour automatic methods to evaluate our ground truth pipeline outputs (stereo-rectified images and disparity maps). The evaluation focuses on the rectification error measure between the left and right stereo-rectified images; and on the accuracy of the disparity maps.

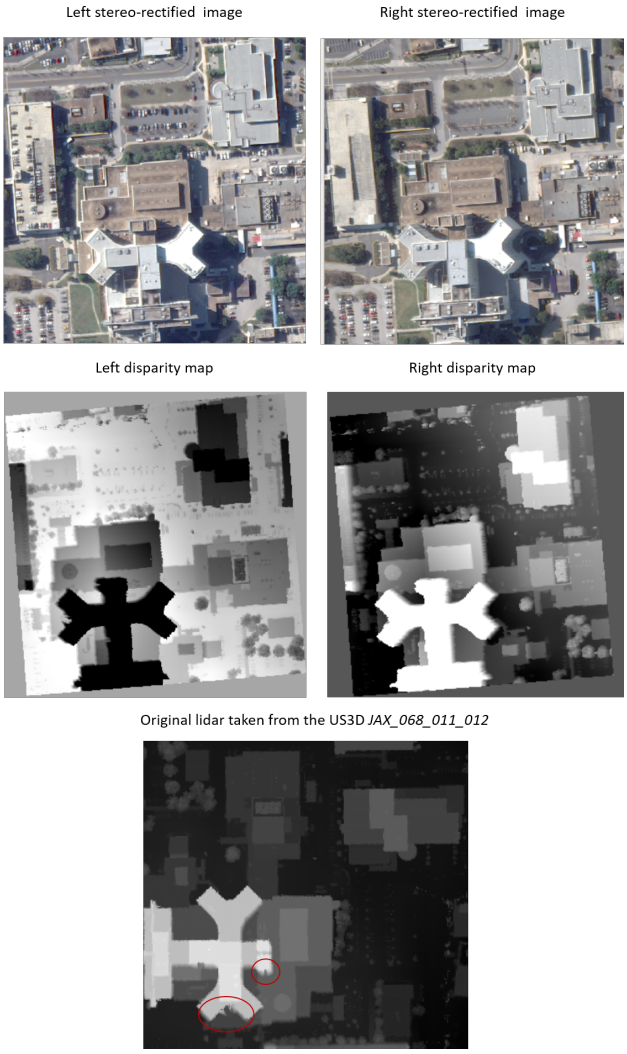

Figure 3. Top 4 images: Outputs of our ground truth pipeline produced from the US3D JAX_068_011_012. Bottom: input lidar raster with a 50cm resolution. It shows our disparity maps dependance on the input lidar raster accuracy (cf red circles added on the lidar, and corresponding areas in the disparity maps); and illustrates the time discrepancy between the inputs (cf cars).

Our assessment is based on 122 Track3 US3D sensor imagery pairs and lidar, as inputs (Le Saux et al., 2019). Track3 US3D sensor imagery is made of DigitalGlobe WorldView-3 images over Jacksonville in Florida, and over Omaha in Nebraska (Bosch et al., 2019). Our pair selection is quasi-random, we aimed at computing the whole Track3 US3D database, but we lacked time for this publication. The stereoscopic angles range of these 122 pairs is really large [0.087; 0.6], that is worst case for our pipeline evaluation. US3D lidar size is around 512×512-pixel with a 50cm resolution. As the lidar areas are small enough, in the following tests, we skip the bundle adjustment step in favour of a SIFT based correction of the right rectifying grid, and a 3D translation of the lidar. In the following tests, we also use a local ratios median instead of the local ratios, assuming that the image resolution and the stereoscopic angle between the views are nearly constants on such small areas (around 853×853-pixel size on WV3 images with a 30cm nadir resolution). Future tests on this dataset, and obviously on wider areas, will be done while using the pixel-wise ratio, presented Eq. 1.

	Pairs nb.	Matches nb.	Mean of the mean rectification errors (pix)	Mean of the mean absolute rectification errors (pix)
SIFT	122	106284	-0.009	0.76
SURF	122	143381	-0.011	0.56

Table 1. Rectification errors (in pixels) computed using SIFT or SURF matches over 122 stereo-rectified images pairs, generated from Track3 US3D sensor imagery pairs and lidar by our ground truth pipeline.

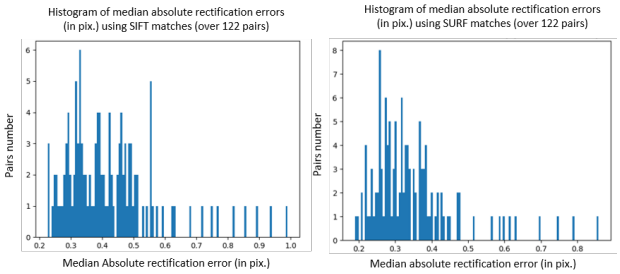

Figure 4. Histograms of median absolute rectification errors (in pixels), using SIFT on the left, and SURF on the right, over the same 122 pairs as in Tab. 1.

4.1 Rectification error

The rectification error corresponds to the row delta between two homologous points in the left and right stereo-rectified images. Using the aboved mentioned 122 Track3 US3D sensor imagery pairs as inputs of our ground truth pipeline, we compute the residual rectification error of our stereo-rectified images, using SIFT and SURF measures. Note that the rectification accuracy presented below is related to these SIFT and SURF matches precision.

We remind that the rectification error is directly computed before and after the right stereo-rectifying grid correction, using a vlfeat implementation of SIFT (cf section 3.2.2).

In this rectification error assessment, we are using another SIFT implementation than the vlfeat one. Because SIFTs are also involved in CARS rectification process, and though we do not share the same implementation, we also present results obtained with SURF matches (Bay et al., 2008). These SIFT and SURF matches implementations are available in the opensource Orfeo ToolBox.

Over the 122 stereo-rectified images pairs, we obtained 106284 SIFT matches and 143381 SURF matches, after outliers rejection step based on the matches distance in the vertical and horizontal directions. The vertical filtering is a 10-pix vertical threshold between two SIFT matches, assuming that the rectification error is less than 10 pixels. The horizontal filtering is deduced from the lidar heights. The SIFT and SURF based statistics were computed independently for each images pair.

According to these SIFT and SURF measures, the rectification error is well centered, cf the global mean of the 122 mean rectification errors in Tab. 1 (-0.009 pixel for SIFT and -0.011 pixel for SURF). Note that we obtain a difference of 35% between the SIFT and SURF estimations of the global mean of the 122 mean absolute rectification errors (0.76 pixel for SIFT vs 0.56 pixel for SURF). With this difference in mind, it appears difficult to give an accurate estimation of the global mean absolute rectification error over the 122 pairs, but we can guess it is subpixel.

Next, we present the histogram of the median absolute rectification errors obtained by SIFT and SURF, on the 122 pairs cf Fig. 4. According to SIFT measures, 79% of the 122 pairs have a median absolute rectification error, that is less than 0.5 pixel. According to SURF measures, 80% of the 122 pairs have a median absolute rectification error, that is less than 0.4 pixel.

4.2 Disparity accuracy

4.2.1 Image warping The first method is based on inverse warping (Scharstein, Szeliski, 2002). We resample the right stereo-rectified image in the left stereo-rectified image geometry, using the left disparity map. Then, we compare the right resampled image to the left stereo-rectified image. The delta observed between the two images is due to radiometric differences between left and right original images, rectification errors, disparity map errors, occlusions and resampling approximations. In our case, the disparity map errors are linked to the disparity map generation method, but also to the time discrepancy between the lidar and the images. We mainly use inverse warping as a qualitative tool. Fig. 5 shows an example of inverse warping.

4.2.2 SIFT matches on ground or buildings class Then, the second method consists in comparing the disparity map, with the horizontal distance between SIFT matches. To deal with time discrepancy between the lidar and the images, we select keypoints belonging to classes with suspected fewer time changes (e.g.: ground, buildings), than the vegetation. Note that Fig. 3 shows that even ground is subjected to changes (due to cars on parkings for instead). To evaluate the left disparity map, we need a disparity measure at integer positions in the left stereo-rectified images. SIFT points are seldom located at integer positions. We decide to select SIFT matches belonging to areas with locally nearly constant disparity, and to apply the disparity of one SIFT keypoint to its nearer integer position in the left stereo-rectified image. That means that we make the assumption that the selected class has a half-pixel piecewise constant disparity (with a pixel size corresponding to the pixel size of the left stereo-rectified image).

In this evaluation, we use the ground and buildings classes of the Track3 US3D classification. This classification seems to have been done on lidar. That means that we still suffer from the time discrepancy between the lidar and the images. This classification has also its own shortcomings (some lidar buildings pixels being classified as ground, or the contrary). But it was available, and useful. Then we project this classification into the stereo-rectified images pairs geometries. Due to these classification approximations, the following class-driven statistics should be considered as preliminary work. Finally we assume that the half-pixel piecewise constant disparity is satisfied on the ground class, and on the buildings class, but it can sometimes be false on both.

Over the 122 stereo-rectified images pairs, we use the same SIFT and SURF keypoints as in section 3.2.2. We just select the ones belonging either to the ground class, or to the buildings class. We also discard the pairs that have less than 33 matches for each pair in the considered class, for statistical representativeness. Regarding SIFT, it gives 77 pairs for the ground class,

	Pairs nb.	Matches nb.	Mean of the mean disparity errors (pix)	Mean of the mean absolute disparity errors (pix)
SIFT Ground	77	8824	0.43	1.26
SIFT Buildings	49	3750	-0.20	0.89
SURF Ground	95	13103	0.13	1.24
SURF Buildings	53	4674	-0.10	0.71

Table 2. Disparity errors (in pixels), computed using SIFT or SURF matches, over stereo-rectified images pairs and disparity maps, extracted from the 122 pairs of Tab. 1, with the 33 matches minimum threshold per class and per keypoints type.

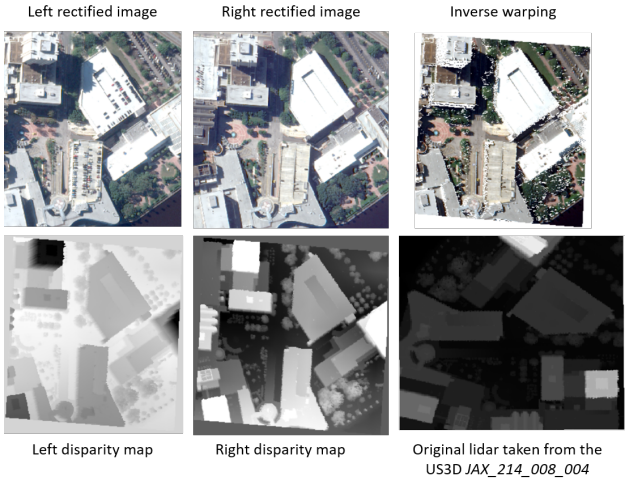

Figure 5. Left and middle: Outputs of our ground truth pipeline produced from the US3D JAX_214_008_004. Right: inverse warping and input lidar raster. This pair obtains the worse mean disparity error on buildings class according to SIFT (0.71 pix). Time discrepancy is visible on roof tops and a little on ground. Top left building in the left disparity map also shows an example of façade reconstruction from the nadir lidar.

and 49 pairs for the buildings class. The statistics were computed independently for each images pair. Here below we focus on the SIFT based results, as the SURF estimations are almost always more favourable than the SIFT ones.

Using SIFT, we first present the global mean of the mean disparity errors (0.43 for ground class, and 0.20 for buildings class); and the global mean of the mean absolute disparity errors (1.26 for ground class, and 0.89 for buildings class), cf Tab. 2. The global disparity accuracy is finally worse on ground class than on buildings class. These ground class statistics could be worsen by classification mistakes (building edges being sometimes classifed as ground). This ground class is also probably more subjected to changes than the buildings class on this urban dataset (cf cars on parkings in Fig. 3, even if there are also parkings on roof tops, such as on Fig. 5). Note that the global mean disparity error is sub-pixel for the buildings class, even for the mean of the mean absolute errors.

Then, we present the histogram of the median absolute disparity errors, obtained using SIFT matches on ground and buildings

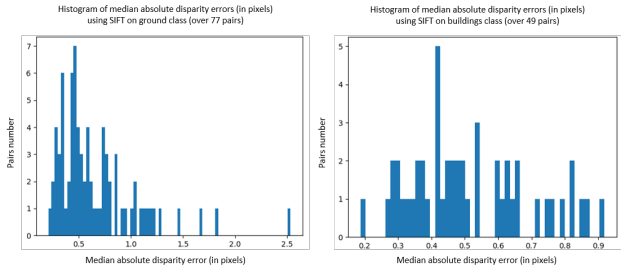

Figure 6. Histograms of median absolute disparity errors (in pixels), computed using SIFT matches, on ground class on the left (77 pairs), and on buildings class on the right (49 pairs). Same pairs as in Tab. 2.

classes, cf Fig. 6. Regarding the ground class, 83% of the 77 pairs have a sub-pixel median absolute disparity error. Regarding the buildings class, all 49 pairs present a sub-pixel median absolute disparity error, with 57% having a median absolute disparity error that is less than 0.5 pixel.

5. GROUND TRUTH USE CASE

In this section, we briefly present the Pandora stereo-matching framework, that will be made publicly available as an open-source software. Then we introduce qualitative results obtained with a Pandora MC-CNN network implementation, trained on a preliminary ground truth dataset.

5.1 Foreword: Pandora framework in a nutshell

Pandora is a new stereo-matching framework, inspired by the work of (Scharstein, Szeliski, 2002). To estimate a disparity map from two stereo-rectified images, Pandora provides the following steps: matching cost computation, cost aggregation, cost optimization, disparity computation, subpixel disparity refinement, disparity filtering and validation.

Pandora is easy to configure. One Pandora configuration can for instance emulate a SGM-like behaviour. Another one can emulate a MC-CNN-like one (Žbontar, LeCun, 2015), while a third one can simply perform standard block-matching (i.e.: a pure local approach without optimization).

Pandora modularity and easy configuration make this software a useful tool for the stereo-matching step of 3D restitution pipelines, as the stereo-matching configuration can be adapted to the area to reconstruct.

Pandora will be the stereo-matching algorithm of the CO3D ground segment. The Pandora framework will be made publicly available on https://github.com/CNES/Pandora.

5.2 Ground truth use case for training

Regarding the ground truth use case, we first aim at training a deep learning network on our ground truth dataset. As a network, we choose the MC-CNN-fast, because an implementation is already available in Pandora. This network computes a local cost volume, that is then optimized in a close variant of the SGM optimization. As a dataset, we choose the 122 stereo-rectified images pairs and disparity maps, introduced and evaluated in section 4. We remind that this ground truth dataset is made of multi-date WorldView-3 images and lidar, over

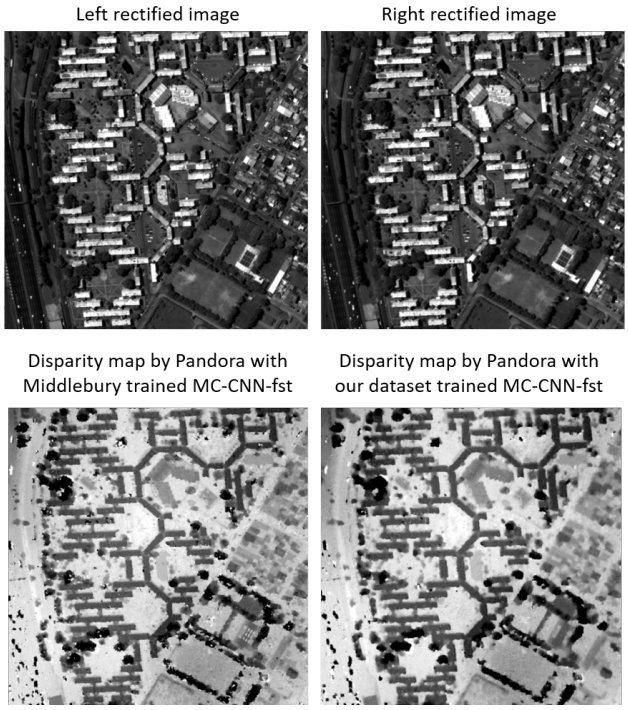

Figure 7. Top: rectified images pair extracted from a WV03 single pass acquisition over Buenos Aires. Bottom: disparity maps computed by the Pandora pipeline, using MC-CNN-fst trained either on Middlebury (left), or on our preliminary ground truth dataset (right).

Figure 8. Top: rectified images pair extracted from a PHR-1A single pass acquisition over London. Bottom: disparity maps computed by the Pandora pipeline, using MC-CNN-fst trained either on Middlebury (left), or on our preliminary ground truth dataset (right).

the USA, without any kind of inputs selection strategy. It implies that this ground truth dataset should be considered as a preliminary one. We train our MC-CNN-fast network implementation on two independent datasets: the Middlebury dataset and our preliminary ground truth dataset. Note that, except the occluded patches (occluded according to the left-right diparity maps cross-checking), we keep all patches extracted from our ground truth dataset. That means that we also keep image patches showing important time differences (cf cars on ground Fig. 3).

Then, we perform a qualitative comparison of the disparity maps produced by the Pandora pipeline, in the MC-CNN configuration, with MC-CNN-fast trained either on Middlebury, or on our preliminary ground truth dataset.

This comparison is done on two urban areas, acquired by two different sensors, both in single pass acquisition. Both inputs pairs are rectified by CARS. The first case input is a WorldView-03 images pair over Buenos Aires, with a stereoscopic angle of 0.12 (cf Fig. 7). It comes from the hexuplet of the IARPA Multi-view Stereo 3D Mapping Challenge (Bosch et al., 2017). The second case input comes from a PleiadesHR-1A images pair over London, with a stereoscopic angle of 0.38 (cf Fig. 8). Note that the Buenos Aires case uses the same sensor as the one used to generate our preliminary ground truth dataset, unlike the second one, that is PleiadesHR based. Both cover a different geographic area than the one used to generate our dataset.

On these two mono-date use cases, both trainings give relatively close qualitative results on the final disparity maps (cf Fig. 7 and Fig. 8). Note that ours succeeds in adapting to new geographical areas and sensor. The training on ours generates a little smoother disparity maps, compared to the Middlebury training. It is visible on buildings edges, on trees, but also on flat areas, such as buildings roofs. This smoothing could be linked to the multi-date images and lidar that we use as inputs of our ground truth pipeline (the Middlebury dataset being without date related changes), and to the lack of patches selection. However, we think that these preliminary results are quite good.

Though easy for us to try on (since Pandora can emulate it), the MC-CNN-fast might not be the best candidate to highlight what a satellite imagery based dataset can offer to neural network approaches. Indeed, Middlebury trained MC-CNN-fast already behaves good on satellite imagery (cf Fig. 8 and 7). MC-CNN uses images patches as inputs, whereas end-to-end networks use the whole left and right stereo-rectified images as inputs. Thanks to this work, we will now be able to train end-to-end networks on satellite imagery based dataset.

Future works will improve our preliminary dataset (working on the pairs selection and on the methodology). We will also train end-to-end stereo-matching networks on this improved dataset.

6. CONCLUSION

Though still in an early stage at CNES, our ground truth pipeline already achieves competitive rectification and disparity accuracies. This preliminary assessment gave a systematic subpixel median absolute disparity error, on the buildings class, using SIFT matches. As a use case, we introduced qualitative results obtained by a deep learning network, trained on a preliminary ground truth dataset. Future works will improve this

preliminary dataset and train end-to-end stereo-matching networks on this improved dataset. Because it also relies on the possibility to directly stereo-rectify wide areas without the need for tiles stitching, we believe this work should be promising.

7. ACKNOWLEDGEMENTS

The authors would like to thank Daniel Greslou for the fruitful discussions. The authors would like to thank the Johns Hopkins University Applied Physics Laboratory and IARPA for providing the data used in this study, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee for organizing the Data Fusion Contest.

REFERENCES

Bay, H., Ess, A., Tuytelaars, T., Gool], L. V., 2008. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, 110(3), 346 - 359.

Bosch, M., Foster, K., Christie, G., Wang, S., Hager, G. D., Brown, M., 2019. Semantic stereo for incidental satellite images. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1524–1532.

Bosch, M., Leichtman, A., Chilcott, D., Goldberg, H., Brown, M., 2017. Metric Evaluation Pipeline for 3D Modeling of Urban Scenes. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-1/W1, 239–246.

d'Angelo, P., Kuschk, G., 2012. Dense multi-view stereo from satellite imagery. *2012 IEEE International Geoscience and Remote Sensing Symposium*, 6944–6947.

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

d'Angelo, P., Cerra, D., Azimi, S. M., Merkle, N., Tian, J., Auer, S., Pato, M., de los Reyes, R., Zhuo, X., Bittner, K., Krauss, T., Reinartz, P., 2019. 3d semantic segmentation from multi-view optical satellite images. *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 5053–5056.

Facciolo, G., De Franchis, C., Meinhardt-Llopis, E., 2017. Automatic 3d reconstruction from multi-date satellite images. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1542–1551.

Hirschmuller, H., 2008. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2), 328-341.

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

Lebegue, L., Cazala-Hourcade, E., Languille, F., Artigues, S., Melet, O., 2020. Co3d, a worldwide one-meter accuracy dem for 2025. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.

Melet, O., Youssefi, D., L'Helguen, C., Michel, J., Sarrazin, E., Languille, F., Lebegue, L., 2020. Co3d mission digital surface model production pipeline. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.

Menze, M., Heipke, C., Geiger, A., 2018. Object Scene Flow. *ISPRS Journal of Photogrammetry and Remote Sensing*, 140, 60 - 76. Geospatial Computer Vision.

Michel, J., Sarrazin, E., Youssefi, D., Cournet, M., Buffe, F., Delvit, J., Emilien, A., Bosman, J., Melet, O., L'Helguen, C., 2020. A new satellite imagery stereo pipeline designed for scalability, robustness and performance. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.

Nuth, C., Kääb, A., 2011. Co-registration and bias corrections of satellite elevation data sets for quantifying glacier thickness change. *The Cryosphere*, 5(1), 271–290.

Patil, S., Comandur, B., Prakash, T., Kak, A. C., 2019. A New Stereo Benchmarking Dataset for Satellite Images. *arXiv e-prints*.

Qin, R., 2016. RPC Stereo Processor (RSP) – A Software Package for Digital Surface Model and Orthophoto Generation from Satellite Stereo Imagery. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, III-1, 77–82.

Rupnik, E., Pierrot-Deseilligny, M., Delorme, A., 2018. 3D reconstruction from multi-view VHR-satellite images in MicMac. *ISPRS Journal of Photogrammetry and Remote Sensing*, 139, 201 - 211.

Scharstein, D., Szeliski, R., 2002. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International Journal of Computer Vision*, 47(1), 7–42.

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 of Photogrammetry and Remote Sensing*, 116, 101 - 117.

Youssefi, D., Michel, J., Sarrazin, E., Buffe, F., Cournet, M., Delvit, J., L'Helguen, C., Melet, O., Emilien, A., Bosman, J., 2020. Cars: A photogrammetry pipeline using dask graphs to construct a global 3d model. *IGARSS - IEEE International Geoscience and Remote Sensing Symposium*.

Žbontar, J., LeCun, Y., 2015. Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. *arXiv e-prints*, arXiv:1510.05970.