IMPROVING SEMI-GLOBAL MATCHING: COST AGGREGATION AND CONFIDENCE MEASURE

Pablo d'Angelo

German Aerospace Center (DLR), Remote Sensing Technology Institute D-82234 Wessling, Germany
email: Pablo.Angelo@dlr.de

Commission I, WG I/4

KEY WORDS: Dense Matching, Digital Elevation Model, Stereo, Benchmark, Accuracy

ABSTRACT:

Digital elevation models are one of the basic products that can be generated from remotely sensed imagery. The Semi Global Matching (SGM) algorithm is a robust and practical algorithm for dense image matching. The connection between SGM and Belief Propagation was recently developed, and based on that improvements such as correction of over-counting the data term, and a new confidence measure have been proposed. Later the MGM algorithm has been proposed, it aims at improving the regularization step of SGM, but has only been evaluated on the Middlebury stereo benchmark so far. This paper evaluates these proposed improvements on the ISPRS satellite stereo benchmark, using a Pleiades Triplet and a Cartosat-1 Stereo pair. The over-counting correction slightly improves matching density, at the expense of adding a few outliers. The MGM cost aggregation shows leads to a slight increase of accuracy.

1. INTRODUCTION

Creation of digital elevation models by automatic image matching of airborne or spaceborne optical data is one of the basic procedures in photogrammetry. While mature and well performing stereo algorithms exist, there is still room for improvements. In the last years, the Semi-Global Matching (SGM) algorithm (Hirschmüller, 2008) has been successfully applied to a variety of stereo problems.

It has proven to be very robust and provides a good compromise between computational speed and matching quality. However, there is still a need for improvements, for example, a pixel-wise reliability score would be very helpful for further processing such as DSM fusion and editing.

Additionally, the regularization performed by the SGM algorithm is not as strong as in other, computationally more demanding global optimization algorithms, such as total variation based algorithms.

2. RECENT SGM IMPROVEMENTS

The cost aggregation algorithm is the core of the SGM method, has been used as a basic component by many stereo algorithms, but itself hasn't been investigated much. However, recently (Drory et al., 2014) analyzed SGM from a theoretical standpoint and derives it as a special case of Belief Propagation. (Facciolo et al., 2015) proposes improvements to the aggregation algorithm. However, their contributions were evaluated on Middlebury close range data (Scharstein et al., 2014), with often has different properties than close range data.

2.1 Basic SGM algorithm

The main components of SGM are matching cost computation and cost aggregation.

The matching cost $C(\mathbf{p}, \mathbf{q})$ computes a similarity value for potentially matching pixels in two images. Using the epipolar geometry, matching costs are computed for all potentially matching pixels in the image pair. For all examples in this paper the Census transform (Zabih and Woodfill, 1994) is used. The window size as set to 7 by 9 pixels. In a thorough evaluation of many matching cost functions (Hirschmüller and Scharstein, 2009), Census turned out to be a very robust and reliable cost function with good performance.

As the matching costs based on single pixel values or small windows are ambiguous, regularization is used to ensure a well behaved reconstruction. For semi-global matching, the matching step is cast into an energy minimization problem. The following, discontinuity preserving energy should be minimized:

$$E(D) = \sum_{\mathbf{p}} (C(\mathbf{p}, D_{\mathbf{p}}) + \sum_{\mathbf{p},\mathbf{q} \in N_{\mathbf{p}}} V(D_{\mathbf{p}}, D_{\mathbf{q}}) \qquad (1)$$

with

$$V(d, d') = \begin{cases} 0 & \text{if } \mathbf{d} = \mathbf{d}' \\ P1 & \text{if } |\mathbf{d} - \mathbf{d}'| = 1 \\ P2 & \text{otherwise} \end{cases} \qquad (2)$$

The function C defines the matching cost between the image pixels for each pixel location $\mathbf{p}$ in the first image and the corresponding pixel in the other image, as defined by the disparity map D. The pairwise term $V(\mathbf{p}, \mathbf{q})$ penalize disparity changes in the neighborhood N_p of each position p. The penalty P_1 is added for all disparity changes equal to one pixel. At larger discontinuities (disparity change > 1 pixel), a fixed cost P_2 is added. This cost function favors similar or slightly changing disparities between neighborhood pixels, and thus stabilizes not only the matching in image areas with weak contrast, but also allows large disparity jumps in areas with high contrast.

Minimizing Eq. 1 for two dimensional neighborhoods N_p is an NP-complete problem, for which no efficient algorithms exist. In SGM, the minimization is performed by aggregating the cost along a path with direction $\mathbf{r}$:

$$\begin{aligned} L_r(\mathbf{p},d) =& C(\mathbf{p},d) \\ &+ min_{d'}(L_r(\mathbf{p}-\mathbf{r},d') + V(d,d')) \\ &- min(L_r(\mathbf{p}-\mathbf{r})) \end{aligned} \quad (3)$$

Summing L for all N_{dir} directions provides the aggregated cost S:

$$S(\mathbf{p},d) = \sum_{\mathbf{r}} L_{\mathbf{r}}(\mathbf{p},d). \quad (4)$$

Usually, 8 directions, vertical, horizontal and diagonals are used, but 16 directions gives better results and reduce streaking artifacts.

The disparity map D is computed by finding the minimum aggregated cost S for each pixel p in the first image. Subpixel accuracy is archived by fitting a local parabola to the aggregated costs around the minimum. Further sub-pixel accuracy can be obtained by sampling the disparity space with 0.5 pixel steps.

Matching is performed from first to the second and second to the first image, and only consistent disparities passing the left-right check are kept. Small, independent disparity segments are identified and removed as outliers. The disparity image is reprojected into a DSM with the desired projection and grid spacing. Support data such as confidence layers are also reprojected. Finally any remaining no-data areas are filled using inverse distance weighted interpolation.

2.2 Over-counting correction

The connection between SGM and Belief Propagation (BP) has been established by (Drory et al., 2014). They show that SGM can be interpreted as the first pass of min-sum BP. Compared to belief propagation, SGM counts the data term C N_{dir} times, thus it should be subtracted when computing S:

$$S(\mathbf{p},d) = \sum_{\mathbf{r}} L_{\mathbf{r}}(\mathbf{p},d) - (N_{dir}-1) * C(\mathbf{p},d). \quad (5)$$

2.3 More global matching

The regularization performed by the SGM algorithm is not as strong as in other, computationally more demanding global optimization algorithms. In contrast to global methods, where all pixels influence each other, SGM performs scanline optimization (cost aggregation) in different aggregation directions, thus each pixel in influenced only by pixels located on 8 horizontal, vertical or diagonal lines. The more global matching (MGM) proposed by (Facciolo et al., 2015) provides a simple extension to improve the regularization by additionally considering the already aggregated previous scanline:

$$\begin{aligned} L_r(\mathbf{p},d) =& C(\mathbf{p},d) \\ &+ \sum_{\mathbf{x}\in\{\mathbf{r},\mathbf{r}_\perp\}} \frac{1}{2} \min_{d'}(L_r(\mathbf{p}-\mathbf{x},d') + V(d,d')) \\ &- min(L_r(\mathbf{p}-\mathbf{r})) \end{aligned} \quad (6)$$

Thus the update is then influenced by the upper left quadrant of $\mathbf{p}$, not just the pixels on path $\mathbf{r}$.

2.4 Matching confidence

Per pixel confidence or even accuracy values would be beneficial for further processing such as data fusion and object extraction. For example the most important information for most DSM fusion algorithms are per pixel weight maps.

When multiple independent stereo pairs are available, for example when computing wide area DSMs (Uttenthaler et al., 2013, Fujisada et al., 2012), values such as number of matches per pixel and their height standard deviation can be computed and used as accuracy values. In the photogrammetric community, the dependency of surface slope on the height error has been used for medium resolution DEMs, but it is unclear if this can be applied to VHR satellite data and detailed DSMs that include discontinuities such as building edges and other small scale features.

Many quality metrics based on data computed during the image matching have been proposed in the past (Hu and Mordohai, 2012). These include correlation coefficient, curvature of the sub-pixel parabola, and others.

For SGM, (Drory et al., 2014) proposed a new uncertainty value which is based on an estimation of the lower bound of the energy values. The lower bound S_m for the aggregated cost at each each pixel $\mathbf{p}$ can be computed as sum of directional minima.

$$S_m(\mathbf{p}) = \sum_{i}^{N_{dir}} min^i_{d_{\mathbf{p}}}(L^i(d^i_{\mathbf{p}}) - \frac{N_{dir}-1}{N_{dir}} C(\mathbf{p},d^i_{\mathbf{p}})) \quad (7)$$

$$A_{lb}(\mathbf{p}) = S(\mathbf{p}) - S_m(\mathbf{p}) \quad (8)$$

$A_{lb]}$ is the difference between estimated lower bound and aggregated cost, and used as confidence measure. Two other confidence measurements evaluated in this work are the local surface slope, A_{slp} and the distance between the first and second minima of the aggregated cost A_{mmn}.

3. EVALUATION

3.1 Dataset description

The data of the ISPRS Commission I WG 4 satellite stereo benchmark dataset (Reinartz et al., 2010) is used to evaluates the reliability measure and the modified cost aggregation.

The test region in Catalonia, near Barcelona has been selected due to the availability of several stereo satellite datasets and a good reference data set provided by the Institut Cartogràfic de Catalunya (ICC).

The evaluation is performed on three test sites, show in Table 1 and Fig. 1.

Test area	Lower left corner	Area type
1. Terrassa (Tr)	417400E 4597300N	City, industrial, residential, hills
2. Vacarisses (Va)	409100E 4601700N	Wooded hills, quarry, waste dump
3. La Mola (Mo)	416400E 4608600N	Steep mountainous terrain, forests

Table 1: Position and properties of the selected test areas. The size of each area is 4 km x 4 km. Coordinates are in UTM Zone 31 North.

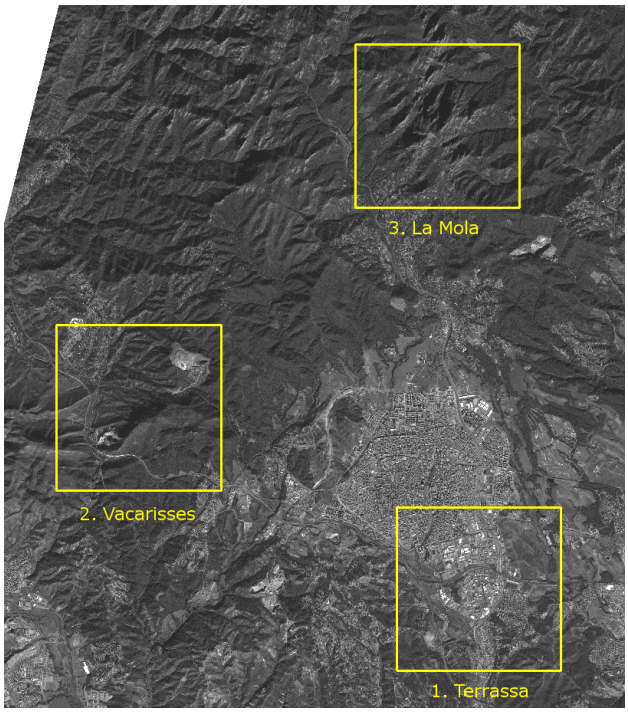


Figure 1: Cartosat-1 image showing the three test areas.

3.1.1 Reference Data The primary reference dataset used in this paper is a 3D point cloud acquired by airborne laser scanning with a density of approximately 0.5 points per square meter, cf. Fig. 2. Only the first pulse returns is used in this study, as the DSM produced by image matching corresponds to the visible surface. The LIDAR data for the Terrassa and Vacarisses test areas was acquired on 26th and 27th November 2007. The LaMola LIDAR data was acquired on 26th November 2007 and 4th May 2008.

3.1.2 Cartosat-1 The test areas are covered by a Cartosat-1 Stereo pair with a ground resolution of 2.5 m and a stereo angle of 31°. Larger shadow areas are visible, as the data has been acquired on the 5th of March 2008. Fig. 1 shows the Cartosat AFT image of the three evaluated test areas.

3.1.3 Pleiades data A Pleiades 1A triplet acquired on 8. January 2013 and provided by Airbus Defense & Space was used as an example for a VHR triplet dataset. The along track viewing angles of the triplet are: $-15.5°$, $-7.5°$ and $16°$. Thus three stereo pairs with convergence angles of $8°$, $23.5°$ and $31.5°$ are possible. This allows reconstruction of fine details in densely build up areas. Due to the winter acquisition, especially the mountainous La Mola area contains deep shadows without usable image content.

3.2 Evaluation procedure

The datasets were bundle adjusted using tie points and ground control provided by the ICC. The generated DSM are thus well registered to the reference data, with a systematic differences in the decimeter range.

All datasets involved in this evaluation where performed using the same basic SGM parameter settings. The used cost function was Census with a 9x7 window, SGM penalties were set to $P1 = 400$ and $P2 = 800$. Images were matched in both directions, and a left and right check is performed. Then the disparity maps are reprojected into a DSM in UTM Zone 31 North, with a grid

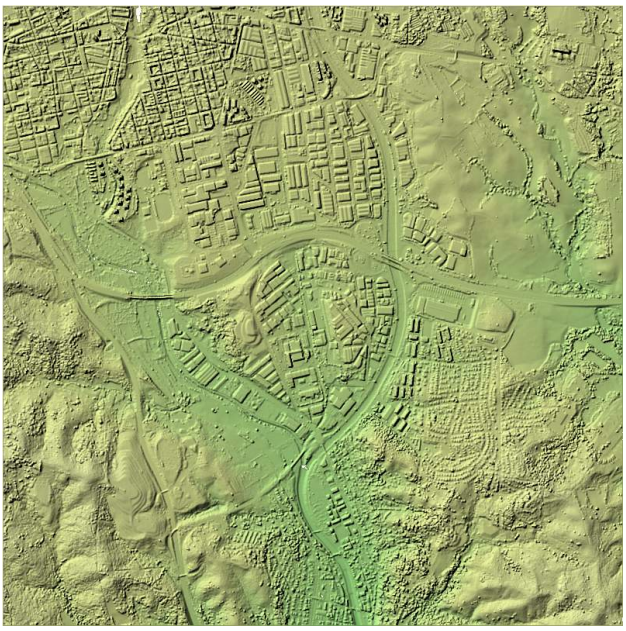


Figure 2: Shaded reference LIDAR DSM of the Terrassa area.

spacing of 1 meter for the Pleiades data, and 5 m for the Cartosat-1 data. For the Pleiades triplet, the 3 possible stereo pairs were matched independently. The pairwise DSMs were averaged to obtain the final DSM. As almost all further processing steps such as orthorectification require a DSM without holes, any remaining no-data values are filled using multi-level B-spline interpolation.

The distance between the LIDAR points and the generated DSM are computed and evaluated statistically. RMSE and normalized median absolute deviation (NMAD) for all results (Höhle and Höhle, 2009). As there is a large time difference between the acquisitions, especially between the Pleiades and the LIDAR data, changed areas, such as new constructed or demolished building, quarries and waste dumps have been manually masked out and were not evaluated. However, there are still some small systematic changes between the LIDAR data due to differences in vegetation.

To evaluate the influence of the over-counting correction and MGM cost aggregation independently, two evaluations were performed.

The over-counting correction was evaluated using the classical SGM algorithm with 16 aggregation directions. Tables 2 and 3 shows that the over-count correction results in a higher completeness of about 1 to 2%. However, slightly worse RMSE, NMAD and bad pixel values are obtained. Other tests with close range imagery showed that the over counting correction lead to a decrease of outliers. It is unclear why the performance on the satellite datasets is different.

The effect of the MGM extension was evaluated on the same dataset, but with 8 aggregation directions only, as our current MGM implementation cannot aggregate along 16 directions. Tables 4 and 5 show the evaluation results.

In general, very similar accuracy values are archived, with slightly better values for MGM8. The results on the Vacarisses area are interesting, here for Cartosat-1, MGM8 has a higher RMSE but lower NMAD than SGM8, whereas for Pleiades, MGM8 performs better on both. A closer evaluation shows that the north face of the hill covered by a dark shadow, and only very noisy image data is available for this region. Here MGM8 does provides

Test area	fix OC	Complete-ness	RMSE	NMAD	BAD
Tr	yes	98.67 %	3.44 m	2.06 m	1.87 %
Tr	no	97.99 %	3.35 m	1.99 m	1.72 %
Va	yes	95.53 %	3.75 m	2.90 m	1.87 %
Va	no	94.17 %	3.67 m	2.77 m	1.80 %
Mo	yes	89.34 %	7.35 m	3.74 m	6.18 %
Mo	no	87.45 %	7.41 m	3.58 m	6.11 %

Table 2: Results for SGM with 16 directions with and without over-counting (OC) correction evaluation on the Cartosat-1 stereo pair. The BAD column gives the percentage of pixels with errors > 10 m.

Test area	fix OC	Complete-ness	RMSE	NMAD	BAD
Tr	yes	96.33 %	2.35 m	0.68 m	4.10 %
Tr	no	95.40 %	2.31 m	0.68 m	3.96 %
Va	yes	92.33 %	3.38 m	1.35 m	8.49 %
Va	no	89.91 %	3.93 m	1.33 m	9.24 %

Table 3: Over-counting evaluation on Pleiades Triplet. The La Mola region was not included, as the very dark shadows without image details in the mountainous regions lead to large no-data regions. The BAD column gives the percentage of pixels with errors > 5 m.

a few small height segments in the shadow area, leading to much better interpolation of the larger no-data area. In this case, it also outperforms SGM16, cf. Tab. 3, which in most other cases performs similar to MGM8. Visual inspection of the DSMs shows that MGM provides slightly denser results, at the cost of also increasing the size of outlier segments. For this evaluation, disparity segments smaller than 10 pixels were removed as outliers. When stronger outlier rejection is used, the number of outlier will be slightly reduced, at the expense of loosing small details, such as the high residential buildings in the Terrassa data. Figure 3 shows the results on a part of the Terrassa area.

3.3 Confidence measures

Confidence or accuracy measures are evaluated by comparison of measure with the LIDAR data. The accuracy of a DSM point depends on various factors, such as image noise, texture, shadow and local surface slope. A direct modeling of the error is thus not possible. Instead, most approaches try to use variables such as local slope or metrics computed by the image matching algorithm as indicators for the confidence.

The height differences between the LIDAR DSM and the MGM8 DSM are evaluated. Often, statistical accuracy measures, such as RMSE, are calculated for several slope classes, and used as indicator for the DSM accuracy. For example, Figures 4, 5, 6 and 7 show the NMAD vs slope, aspect, A_{mmn} and A_{lb}. The behavior of the Cartosat-1 and Pleiades DSM show similar trends. Due to the large shadow areas, the Vacarisses and La Mola areas show higher errors for northern aspects.

Test area	Method	Complete-ness	RMSE	NMAD	BAD
Tr	MGM8	98.28 %	3.39 m	2.02 m	1.81 %
Tr	SGM8	98.62 %	3.43 m	2.07 m	1.85 %
Va	MGM8	95.51 %	4.11 m	2.83 m	1.82 %
Va	SGM8	95.05 %	3.79 m	2.91 m	1.98 %
Mo	MGM8	92.79 %	6.65 m	3.60 m	5.39 %
Mo	SGM8	92.57 %	6.78 m	3.68 m	5.77 %

Table 4: SGM8 vs MGM8 evaluation on Cartosat-1 stereo pair. The BAD column gives the percentage of pixels with errors > 10 m.

Test area	Method	Complete-ness	RMSE	NMAD	BAD
Tr	MGM8	95.99	2.31	0.68	3.92%
Tr	SGM8	96.27	2.35	0.68	4.09%
Va	MGM8	92.40	3.01	1.31	7.70%
Va	SGM8	92.16	3.56	1.36	8.67%

Table 5: SGM8 vs MGM8 evaluation on Pleiades Triplet. The BAD column gives the percentage of pixels with errors > 5 m.

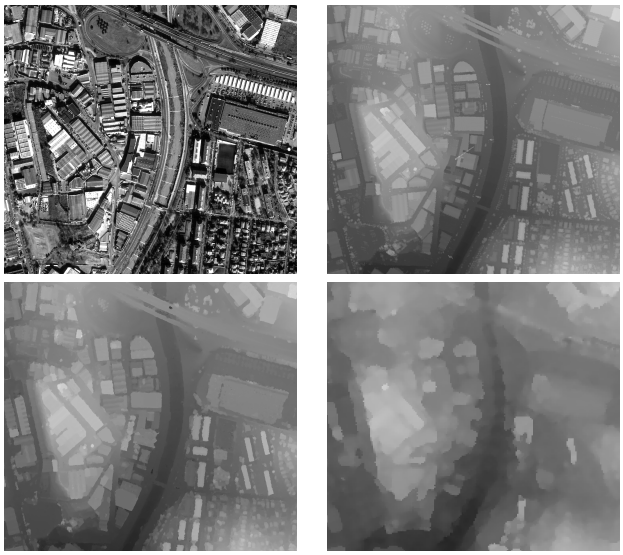


Figure 3: Detail of MGM8 results on the Terrassa area. Top row: Pleiades ortho image, LIDAR reference. Bottom Row: Pleiades DSM, Cartosat-1 DSM.

The height error values are often not normally distributed, and cannot be completely described by a single number. Instead, joint histograms of the height error and confidence value show the full distribution of the errors. For example, Figures 4, 6 and 7 show similar trends, but the underlying distributions in Fig. 8 look very different.

Figure 8 visualizes that the height error vs confidence variable distribution of the Pleiades Terrassa triplet. It can be seen that the distribution is centered and quite symmetric. A narrow distribution at low slopes is expected, which should progressively widens as the slope gets higher. However, there is a tendency that higher slopes show less variation in height error than low slopes. When using the aggregated cost distance A_{mmn}, it is visible that the small distances cause a larger variation, thus it is a better error indicator than the slope. The confidence A_{lb} is proposed in (Drory et al., 2014). For $A_{lb} = 0$ all aggregation paths agree on a single disparity, indicating a confident solution. However, for large A_{lb} values above 1000, the height error decreases again. A_{lb} is thus does not provide a strong indication for height errors, here A_{mmn} performs better.

4. CONCLUSION

This paper investigates the performance of several recently proposed improvements for Semi-Global Matching in the context of

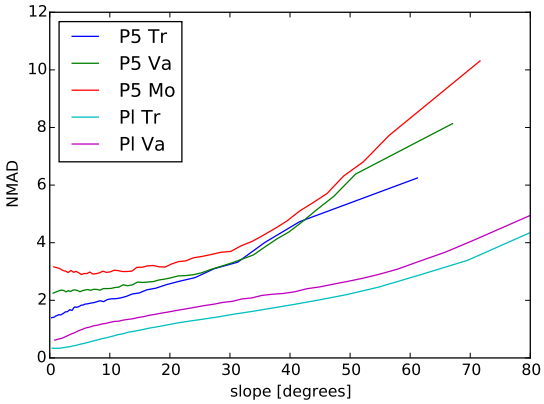


Figure 4: Dependence of height error on slope for Cartosat-1 (P5) and Pleiades (Pl) DSMs.

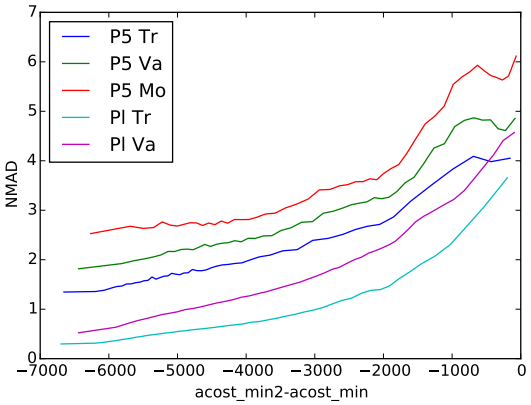


Figure 6: Dependence of height error on difference of first and second aggregated cost minima A_{mmn}.

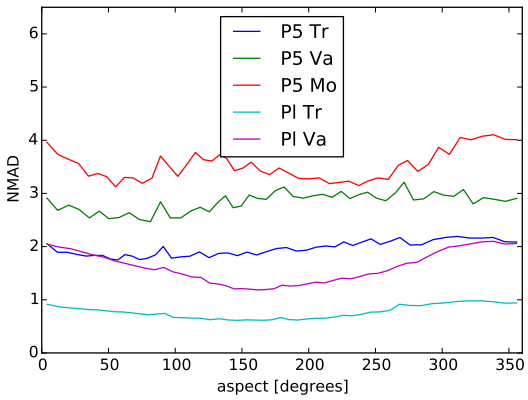


Figure 5: Dependence of height error on aspect.

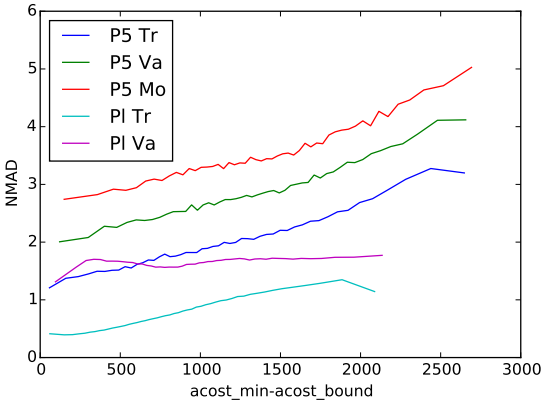
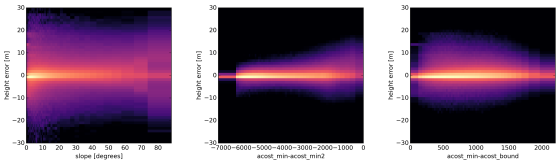


Figure 7: Dependence of height error based on estimated lower bound of energy A_{lb}.



(a) Slope (b) A_{mmn} (c) A_{lb}

Figure 8: Height error vs. confidence variable joint histograms for the Pleiades Terrassa triplet. Color is proportional to the logarithm of density of each joint histogram bin.

DSM generation from satellite data. First the over-counting correction and the confidence measure proposed by (Drory et al., 2014), and the More Global Matching modification by (Facciolo et al., 2015). Using the LIDAR ground truth data, it was found that the over-counting correction results in minimally denser results but also minimal loss in accuracy. The MGM aggregation using 8 directions leads to slightly improved results over 8 directional SGM aggregation. Further investigation, for example with other land-cover types should be performed to check if MGM could increase the performance more significantly in areas with little texture. The uncertainty measure proposed by (Drory et al., 2014) did not perform better than existing methods, such as the energy difference between first and second minima.

Future work could include an extension of the MGM algorithm for aggregation from multiple directions at the same time, as well as a more principled evaluation of confidence functions (Hu and Mordohai, 2012).

5. ACKNOWLEDGMENTS

The authors would like to thank the data providers for their generous support, namely: GAF AG for the Cartosat-1 data, Airbus Defense and Space for the Pleiades triplet and ICC Catalunya for the reference data.

REFERENCES

Drory, A., Haubold, C., Avidan, S. and Hamprecht, F. A., 2014. Semi-global matching: A principled derivation in terms of message passing. In: 36th German Conference on Pattern Recognition.

Facciolo, G., de Franchis, C. and Meinhardt, E., 2015. Mgm: A significantly more global matching for stereovision. In: Proceedings of the British Machine Vision Conference (BMVC), BMVA Press, pp. 90.1–90.12.

Fujisada, H., Urai, M. and Iwasaki, A., 2012. Technical methodology for aster global dem. IEEE Transactions on Geoscience and Remote Sensing 50(10), pp. 3725–3736.

Hirschmüller, H., 2008. Stereo processing by semi-global match-

ing and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(2), pp. 328 – 341.

Hirschmüller, H. and Scharstein, D., 2009. Evaluation of stereo matching costs on image with radiometric differences. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(9), pp. 1582–1599.

Höhle, J. and Höhle, M., 2009. Accuracy assessment of digital elevation models by means of robust statistical methods. ISPRS Journal of Photogrammetry and Remote Sensing 64(4), pp. 398 – 406.

Hu, X. and Mordohai, P., 2012. A quantitative evaluation of confidence measures for stereo vision. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(11), pp. 2121–2133.

Reinartz, P., d'Angelo, P., Krauß, T., Poli, D., Jacobsen, K. and Buyuksalih, G., 2010. Benchmarking and quality analysis of dem generated from high and very high resolution optical stereo satellite data. ISPRS.

Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nesic, N., Wang, X. and Westling, P., 2014. High-resolution stereo datasets with subpixel-accurate ground truth. In: 36th German Conference on Pattern Recognition.

Uttenthaler, A., Barner, F., Hass, T., Makiola, J., d'Angelo, P., Reinartz, P., Carl, S. and Steiner, K., 2013. Euro-maps 3d–a transnational, high-resolution digital surface model for europe. In: ESA Special Publication, Vol. 722, p. 271.

Zabih, R. and Woodfill, J., 1994. Non-parametric local transforms for computing visual correspondence. In: ECCV '94: Proceedings of the Third European Conference-Volume II on Computer Vision, Springer-Verlag, London, UK, pp. 151–158.