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 1/4

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 $C(p,q)$ 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:

$$E(D) = \sum_p C(p, D_p) + \sum_{p,q \in N_p} V(D_p, D_q)$$

(1)

with

$$V(d, d') = \begin{cases} 0 & \text{if } d = d' \\ P_1 & \text{if } |d - d'| = 1 \\ P_2 & \text{otherwise} \end{cases}$$

(2)

The function $C$ defines the matching cost between the image pixels for each pixel location $p$ in the first image and the corresponding pixel in the other image, as defined by the disparity map $D$. The pairwise term $V(p, q)$ 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 $r$:...
\[ L_r(p, d) = C(p, d) + \min_{d'}(L_r(p - r, d') + V(d, d')) \]
\[ - \min(L_r(p - r)) \]

Summing \( L \) for all \( N_{dir} \) directions provides the aggregated cost \( S \):
\[ S(p, d) = \sum_r L_r(p, d). \]

Usually, 8 directions, vertical, horizontal and diagonals are used, but 16 directions give 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 \( S \):
\[ S(p, d) = \sum_r L_r(p, d) - (N_{dir} - 1) \cdot C(p, d). \]

### 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:
\[ L_r(p, d) = C(p, d) + \frac{1}{2} \min_{d'}(L_r(p - r, d') + V(d, d')) \]
\[ - \min(L_r(p - r)) \]

The update is then influenced by the upper left quadrant of \( p \), not just the pixels on path \( r \).

### 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 (Udenthaler 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 pixel \( p \) can be computed as sum of directional minima:
\[ S_m(p) = \sum_{i=1}^{N_{dir}} \min_{d'}^i L^i(p, d'_p) - \sum_{i=1}^{N_{dir}} \frac{1}{N_{dir}} C(p, d'_p) \]

\[ A_{lb}(p) = S(p) - S_m(p) \]

\( A_{lb} \) 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 \( A_{min} \).

### 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 evaluate 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.

<table>
<thead>
<tr>
<th>Test area</th>
<th>Lower left corner</th>
<th>Area type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Terrassa (Tr)</td>
<td>417400E 4597300N</td>
<td>City, industrial, residential, hills</td>
</tr>
<tr>
<td>2. Vacarisses (Va)</td>
<td>409100E 4601700N</td>
<td>Woodyed hills, quarry, waste dump</td>
</tr>
<tr>
<td>3. La Mola (Mo)</td>
<td>416400E 4608600N</td>
<td>Steep mountainous terrain, forests</td>
</tr>
</tbody>
</table>

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.

This contribution has been peer-reviewed.  
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 \( P_1 = 400 \) and \( P_2 = 800 \). 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 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 is 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 MGMS. The results on the Vacarisses area are interesting, here for Cartosat-1, MGMS has a higher RMSE but lower NMAD than SGM8, whereas for Pleiades, MGMS 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 MGMS does provides...
The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLI-B1, 2016

XXIII ISPRS Congress, 12–19 July 2016, Prague, Czech Republic

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.

<table>
<thead>
<tr>
<th>Test area</th>
<th>fix OC</th>
<th>Complete-ness (%)</th>
<th>RMSE (m)</th>
<th>NMAD (m)</th>
<th>BAD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tr</td>
<td>yes</td>
<td>98.67 %</td>
<td>3.44 m</td>
<td>2.06 m</td>
<td>1.87 %</td>
</tr>
<tr>
<td>Tr</td>
<td>no</td>
<td>97.99 %</td>
<td>3.35 m</td>
<td>1.99 m</td>
<td>1.72 %</td>
</tr>
<tr>
<td>Va</td>
<td>yes</td>
<td>95.53 %</td>
<td>3.75 m</td>
<td>2.90 m</td>
<td>1.87 %</td>
</tr>
<tr>
<td>Va</td>
<td>no</td>
<td>94.17 %</td>
<td>3.67 m</td>
<td>2.77 m</td>
<td>1.80 %</td>
</tr>
<tr>
<td>Mo</td>
<td>yes</td>
<td>89.34 %</td>
<td>7.35 m</td>
<td>3.74 m</td>
<td>6.18 %</td>
</tr>
<tr>
<td>Mo</td>
<td>no</td>
<td>84.85 %</td>
<td>7.41 m</td>
<td>3.58 m</td>
<td>11.18 %</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Test area</th>
<th>fix OC</th>
<th>Complete-ness (%)</th>
<th>RMSE (m)</th>
<th>NMAD (m)</th>
<th>BAD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tr</td>
<td>yes</td>
<td>96.33 %</td>
<td>2.35 m</td>
<td>0.68 m</td>
<td>4.10 %</td>
</tr>
<tr>
<td>Tr</td>
<td>no</td>
<td>95.40 %</td>
<td>2.31 m</td>
<td>0.68 m</td>
<td>3.96 %</td>
</tr>
<tr>
<td>Va</td>
<td>yes</td>
<td>92.33 %</td>
<td>3.38 m</td>
<td>1.35 m</td>
<td>8.49 %</td>
</tr>
<tr>
<td>Va</td>
<td>no</td>
<td>89.91 %</td>
<td>3.93 m</td>
<td>1.33 m</td>
<td>9.24 %</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Test area</th>
<th>fix OC</th>
<th>Complete-ness (%)</th>
<th>RMSE (m)</th>
<th>NMAD (m)</th>
<th>BAD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tr</td>
<td>yes</td>
<td>98.28 %</td>
<td>3.39 m</td>
<td>2.02 m</td>
<td>1.81 %</td>
</tr>
<tr>
<td>Tr</td>
<td>no</td>
<td>98.62 %</td>
<td>3.43 m</td>
<td>2.07 m</td>
<td>1.85 %</td>
</tr>
<tr>
<td>Va</td>
<td>yes</td>
<td>95.51 %</td>
<td>4.11 m</td>
<td>2.83 m</td>
<td>1.82 %</td>
</tr>
<tr>
<td>Va</td>
<td>no</td>
<td>95.05 %</td>
<td>3.79 m</td>
<td>2.91 m</td>
<td>1.98 %</td>
</tr>
<tr>
<td>Mo</td>
<td>yes</td>
<td>92.79 %</td>
<td>6.65 m</td>
<td>3.60 m</td>
<td>5.39 %</td>
</tr>
<tr>
<td>Mo</td>
<td>no</td>
<td>92.57 %</td>
<td>6.78 m</td>
<td>3.68 m</td>
<td>5.77 %</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Test area</th>
<th>Method</th>
<th>Complete-ness (%)</th>
<th>RMSE (m)</th>
<th>NMAD (m)</th>
<th>BAD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tr</td>
<td>MGM8</td>
<td>95.99 %</td>
<td>2.31 m</td>
<td>0.68 m</td>
<td>3.92 %</td>
</tr>
<tr>
<td>Tr</td>
<td>SGM8</td>
<td>96.27</td>
<td>2.35 m</td>
<td>0.68 m</td>
<td>4.09 %</td>
</tr>
<tr>
<td>Va</td>
<td>MGM8</td>
<td>92.40</td>
<td>3.01 m</td>
<td>1.31 m</td>
<td>7.70 %</td>
</tr>
<tr>
<td>Va</td>
<td>SGM8</td>
<td>91.19</td>
<td>3.56 m</td>
<td>1.36 m</td>
<td>8.67 %</td>
</tr>
</tbody>
</table>

3.3 Confidence measures

Confidence or accuracy measures are evaluated by comparison of measures 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 show the NMAD vs slope, aspect, \( A_{100} \) and \( A_{150} \) and 7 show the NMAD vs slope, aspect, \( A_{100} \) and \( A_{150} \). 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.

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 \( A_{100} \) is proposed in (Drory et al., 2014). For \( A_{100} = 0 \) all aggregation paths agree on a single disparity, indicating a confident solution. However, for large \( A_{100} \) values above 1000, the height error decreases again. \( A_{100} \) 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...
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


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


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.

