ASSESSING THE ACCURACY OF DENSE IMAGE MATCHING

(or Benchmarking DIM)

Camillo Ressl
camillo.ressl@geo.tuwien.ac.at

Department of Geodesy and Geoinformation
Vienna University of Technology
www.geo.tuwien.ac.at
**Content**

- **Image Matching**: Finding corresponding pixels in $\geq 2$ images with given orientation

  ![Image 1](image1.png) ![Image 2](image2.png) ![DSM](dsm.png)

  Vaihingen, DMC, 8cm, #71+73

  **spatial intersection**

- **Dense**: match every pixel
- **Result**: point cloud (or 2.5D model) of object
Semi-Global Matching (SGM)

- In Computer Vision (CV): Hirschmüller, 2008

Stereo Processing by Semiglobal Matching and Mutual Information

Heiko Hirschmüller

Abstract—This paper describes the Semiglobal Matching (SGM) stereo method. It uses a pixelwise, Mutual Information (MI)-based matching cost for compensating radiometric differences of input images. Pixelwise matching is supported by a smoothness constraint that is usually expressed as a global cost function. SGM performs a fast approximation by pathwise optimizations from all directions. The discussion also addresses occlusion detection, subpixel refinement, and multibaseline matching. Additionally, postprocessing steps for removing outliers, recovering from specific problems of structured environments, and the interpolation of gaps are presented. Finally, strategies for processing almost arbitrarily large images and fusion of disparity images using orthographic projection are proposed. A comparison on standard stereo images shows that SGM is among the currently top-ranked algorithms and is best, if subpixel accuracy is considered. The complexity is linear to the number of pixels and disparity range, which results in a runtime of just 1-2 seconds on typical test images. An in depth evaluation of the MI-based matching cost demonstrates a tolerance against a wide range of radiometric transformations. Finally, examples of reconstructions from huge aerial frame and pushbroom images demonstrate that the presented ideas are working well on practical problems.

- In Photogrammetry: 2011 first implementations appeared: Match-T, Sure
Benchmarks

- Benchmarks are the processes and the results of assessing performance.
- For practice:
  - What accuracies can be achieved?
  - Which parameters influence the accuracy?
- For developers and researchers:
  - Which parameters (cost function, minimization method, …) perform best on various scenarios?
  - SGM triggered by benchmarks performed in CV since 2001: Quote from Hirschmüller, 2008: “Almost all of the currently top-ranked algorithms […] optimize a global energy function.”
Middlebury

- [http://vision.middlebury.edu/stereo/](http://vision.middlebury.edu/stereo/) → image pair
  training + evalutation images, upload result

- **2001**: (0.2 MP)
  - scenes with planar objects
  - Ground truth (GT) disparities labeled by hand

- **2003-2006**: (1.5 MP)
  - 3D objects
  - GT by structured light projector (for coding and intersecting)

- **2014**: (6 MP)
  - like 2003-2006
  - multiple ambient illuminations, complexer scenes
  - GT as sub-pixel disparities

Middlebury

- Much care in preparing the benchmark data (2014)

Scharstein et al., 2014, High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth, GCPR 2014, LNCS 8753
Middlebury (2 views)

- Quality measures:
  - RMS of disparity differences $\Delta D$ (per view)
  - Number of bad pixels: $|\Delta D| > 1$ pix

- Analysis in regions:
  - Textureless regions
  - Occluded regions
  - Depth discontinuity regions
  - (non of the above)

- Overall performance measured by #bad_pixels in non occluded regions (best < 1 %)

- RMS not robust (effected by bad pixels)
  - 2002: 0.05 pix (planar AOI)


Middlebury (multi view)

- [Link](http://vision.middlebury.edu/mview)
- ~ 300 images by robot arm (GSD 0.25 mm)
- 4 different objects
- GT using laser strip scanner; alignment with images using ICP and maximizing photo-consistency
- Participants: 84 (2018)

Seitz, Curless, Diebel, Scharstein, Szeliski, 2006, A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms, CVPR '06

Ground Truth (GT)
### Middlebury (multi view)

- **Quality measures:**
  - **Accuracy** = distance between points in $R$ to closest point in $G$ (best ~ 1-2 GSD, 90%)
  - **Completeness** = distance between points in $G$ to closest point in $R$

Seitz et al., 2006, A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms, CVPR ’06

**Ground Truth (G)**

**Reconstruction to be tested (R)**

**Accuracy:**
- Holes in $G$ are filled ($G'$)
- Closest points to $G'$ are removed from accuracy statistics

**Completeness:**
- Points in $G$ close to border of $R$ or too far away are treated as not covered in $R$
Shortcomings

Background of cited authors: vision-based Driver Assistance Systems, must be accurate on every road, under all kinds of weather conditions, and in any traffic context

- “synthetic (i.e., computer generated stereo pairs) or engineered (i.e., images captured under highly controlled conditions, using structured light for generating ground truth) data do have their own characteristics, and do not cover the “challenges” as occurring in real-world data.”

- “Preliminary experiments show that methods ranking high on established benchmarks such as Middlebury perform below average when being moved outside the laboratory to the real world.”
  [http://www.cvlibs.net/datasets/kitti]

→ evaluate in the real-world
Multi-View Outdoor

- [http://icwww.epfl.ch/multiview/denseMVS.html](http://icwww.epfl.ch/multiview/denseMVS.html)
- Three architectural objects; 8-30 images (6 MP, 3mm GSD)
- Ground Truth (GT) by TLS
- consider the **STD of GT**! (~ 1.3 mm), (motivated by comparing **performance** of TLS and PHO)

- Quality measures:
  - Images are evaluated relative to GT-STD using reference **depth maps** per image.
  - Mean relative error: 2 to 3 * GT-STD

- Participants: 12 (2009), images still available
- Data sets for shape-from-X
- \( X = \) stereo, motion, silhouette, shading, …
- Images using robot arm
- Ground Truth (GT) by laser scanner

- Quality measures:
  - **Accuracy** = RMS of depth map differences \( \Delta D \) (per view)
  - **Completeness** = Number of good pixels: \( |\Delta D| < \delta \) [Digital Numbers]

- Analysis in regions
  - Textureless, Occluded, Depth discontinuity

- Researchers could upload results to web server, but no longer existent?

Motivated by showcasing the properties of various digital aerial cameras wrt image orientation, DEM extraction, radiometry, stereo restitution

- DMC, UCX, ADS40, … with GSD: 8 cm, 20 cm
- ALS (~3 pts/m²) not GT, but participating sensor
- Ground Truth: GPS points, planar objects
- Participants: ~3 (DEM extraction): Match-T, NGATE, SAT-PP
- Data still available? Contact DGPF

Vaihingen/Enz, Germany

East-West ~ 8 km
DGPF 2009

- Quality Measures:
  - GPS points vs. matched DEM, RMS = 0.4 - 1 GSD
  - soccer field: STD(robust) of matched points to common plane, STD = 0.3 - 2 GSD
  - profiles

Our first DIM assessments (2011)

- Investigating the benefit of multi-image matching (using Match-T)

Vienna, UCXp, 6cm GSD (80% / 80%),

Height accuracy: \( \sigma_{\text{MAD}} \approx 4.5 \text{ cm} \) (wrt ALS DSM)

<table>
<thead>
<tr>
<th>pair(s)</th>
<th>n</th>
<th>( \sigma_{\text{MAD}} ) [cm]</th>
</tr>
</thead>
<tbody>
<tr>
<td>80%</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>60%</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>40%</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>80% Fusion</td>
<td>4</td>
<td>8.4</td>
</tr>
<tr>
<td>60% Fusion</td>
<td>3</td>
<td>7.1</td>
</tr>
<tr>
<td>80%+60% Fusion</td>
<td>7</td>
<td>6.6</td>
</tr>
<tr>
<td>80%+60% all strips + cross mods</td>
<td>&lt;70</td>
<td>4.5</td>
</tr>
</tbody>
</table>

In cooperation with Stadt Wien, MA41.
**nota bene:** Use robust statistics, ... but take care!

Example: $dz =$ distance of ALS points (last echo) to their DTM.

What is accuracy of last echos?

**all $dz$ values**

- Mean = 2.8 m
- $\sigma = 6.9$ m
- Median = 1.2 cm
- $\sigma_{\text{MAD}} = 5.0$ cm

$\sigma_{\text{MAD}} = 1.4826$ MAD

$\text{MAD}(x) := \text{median of absolute distances to median}(x)$

$\Rightarrow \quad \sigma_{\text{MAD}}$ only applies if distribution is Gaussian

Other authors refer to $\sigma_{\text{MAD}}$ as $\text{NMAD}$; e.g. Höhle and Höhle, 2009. Accuracy assessment of digital elevation models by means of robust statistical methods. ISPRS Journal 64.

$|dz| < 3 \cdot \sigma$

- Mean = 1.3 m
- $\sigma = 3.7$ m
- Median = 0.9 cm
- $\sigma_{\text{MAD}} = 4.3$ cm

$\Rightarrow \quad \text{Do not use the 3 * sigma rule, if you do not know sigma (and expectation)}$!

Chebyshev's inequality:

$P(|x - \mu| \geq k \cdot \sigma) < 1/k^2$ holds for any distribution!

$k = 3 \Rightarrow P = 11\%$

E.g. if you have 100 values with 21 outliers and you apply the 3*sigma rule (with sigma estimated from this corrupted sample), then afterwards you will have still at least 10 of these outliers.
**nota bene:** Use robust statistics, … but take care!

Example: $dz = \text{distance of ALS points (last echo) to their DTM.}$

What is accuracy of last echos?

all $dz$ values

Mean $= 2.8 \text{ m}$

$\sigma = 6.9 \text{ m}$

Median $= 1.2 \text{ cm}$

$\sigma_{\text{MAD}} = 5.0 \text{ cm}$

$\sigma_{\text{MAD}} = 1.4826 \text{ MAD}$

$\text{MAD}(x) := \text{median of absolute distances to median}(x)$

$|dz| < 4 \times \sigma_{\text{MAD}}$

→ $\sigma_{\text{MAD}}$ only applies if distribution is Gaussian

Histogram still not Gaussian!

In publications, please, report Mean, $\sigma$, and Median, $\sigma_{\text{MAD}}$ (and maybe some quantiles); adding the histogram would be excellent!
Matching problems in shadow areas

ortho-photo, GSD=6 cm

shading (MatchT, SGM, fusion)

standard deviation of fusion

\[ dZ(\text{max-min}): 21\text{cm (sun)} \text{ vs. } 87\text{cm (shadow)} \]

\[ dZ(\text{std}): 4\text{cm (sun)} \text{ vs. } 11\text{cm (shadow)} \]

→ should smoothing/regularization be based on scene content?
Problems at homogenous texture and corners

Shading (MatchT, SGM, fusion)
Matching of trees

a: conifer
b: deciduous tree (leafless)

Z-coding

shading (MatchT, SGM, fusion)

aerial image

In cooperation with Stadt Wien, MA41.
Reconstruction of isolated objects from Satellite

- Only 2-3 images (Pleijades, GSD 0.5 m)
- Reconstruct by matching these objects (Match-T):

Height comparison with manual reconstruction: object size $\geq 15$ pix $\rightarrow$ Z correct $>50\%$
Effect of Land Cover on Height Accuracy

- UCX: GSD 6 cm, (80% / 70%)
- ALS: 4 pts/m² (flown simultaneously)
- Match-T, Sure


<table>
<thead>
<tr>
<th>[cm]</th>
<th>Sealed</th>
<th>Vegetation</th>
<th>Vegetation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>med</td>
<td>std</td>
<td>med</td>
</tr>
<tr>
<td>ALS (last echos)</td>
<td>2.6</td>
<td>5.5</td>
<td>1.1</td>
</tr>
<tr>
<td>DIM</td>
<td>5.4</td>
<td>3.7</td>
<td>19.7</td>
</tr>
</tbody>
</table>

→ DIM over sealed areas better than ALS (bad SNR)
→ DIM at top of grass, ALS penetrates
- Software tested: Sure, MicMac, PMVS, Photoscan
- 8 different objects: GSD = 0.06 mm to 12 cm
- Evaluation of point clouds (not meshes)
- Ground Truth: TLS
- Quality Measures:
  - Flatness: STD = 0.5 - 1 GSD
  - Comparison with TLS-mesh: STD = 0.5 - 1 GSD
  - Profiles
- Problems spotted:
  - Shadows
  - Small structures
  - Sharp discontinuities
Remondino et al., 2014

- Problems at discontinuities
Oblique airborne benchmark
- Zürich: 27 (*5) images with Leica RCD30 Oblique Penta
- GSD: 6 – 13 cm
- Ground Truth: TLS
- Tested software: Sure, Photoscan
- Evaluated: point clouds on facades
- Quality Measures:
  - Density
  - RMS of flatness
  - DIM vs. TLS
  - Profiles
- Data still available: [http://www.ifp.uni-stuttgart.de/ISPRS-EuroSDR/ImageMatching](http://www.ifp.uni-stuttgart.de/ISPRS-EuroSDR/ImageMatching)
- More data at: [http://www2.isprs.org/commissions/comm1/icwg15b/benchmark_main.html](http://www2.isprs.org/commissions/comm1/icwg15b/benchmark_main.html)
EuroSDR & ISPRS WG III/1 2012 - 2016

- Quality Measures:
  - Density
  - RMS of flatness error
  - DIM vs. TLS
  - Profiles

<table>
<thead>
<tr>
<th></th>
<th>GSD [cm]</th>
<th>Density [Points/m²]</th>
<th>RMSE DIM [px]</th>
<th>RMSE DIM-TLS [px]</th>
</tr>
</thead>
<tbody>
<tr>
<td>C S</td>
<td>7.7</td>
<td>120</td>
<td>1.56</td>
<td>1.92</td>
</tr>
<tr>
<td>C PS</td>
<td>7.7</td>
<td>172</td>
<td>2.06</td>
<td>1.77</td>
</tr>
<tr>
<td>R S</td>
<td>7.8</td>
<td>120</td>
<td>1.61</td>
<td>1.53</td>
</tr>
<tr>
<td>R PS</td>
<td>7.8</td>
<td>163</td>
<td>2.12</td>
<td>1.87</td>
</tr>
</tbody>
</table>

RMS (oblique) < 2GSD

Note: Number of matched images was fixed to 5. Thus the full set of overlapping images was not fully exploited, which may explain the larger RMS value. Other authors report RMS of 1 GSD for oblique images (exploiting the full overlap): Zhang, Gerke, Vosselman, Yang, 2018. A patch-based method for the evaluation of dense image matching quality, Int J Appl Earth Obs Geoinformation 70.
Summary

- **Matching nadir images**: STD < 1 GSD
  - Far from 0.1 pix obtained for bundle block because of sub-optimal texture.
  - Be sure to fully exploit the high image overlaps.
  - Dependency on slope (+0.5 GSD over 40° slope) [Müller, Gärtner-Roer, Thee, Ginzler, 2014. Accuracy assessment of airborne photogrammetrically derived high-resolution digital elevation models in a high mountain environment, ISPRS J., 98]

- **Matching oblique images**: STD ~ 1 - 2 GSD

- **Special challenges with oblique**:
  - Illumination changes
  - More occlusions
  - Larger depth of field → GSD variations
  - Full 3D (not 2.5D) workflow required
Conclusions

- Publically available benchmark data sets in Photogrammetry community are a bit short-lived
  → petition: keep data available
- Computer Vision shifts towards real word scenes, maybe Photogrammetry should look for controlled environments?
- Depth accuracy is well handled for nadir (oblique still open research)

- Future:
  Completely open: Evaluate the self-evaluation of DIM
  Focus on noted problem cases
  • Shadows (or homogenous texture in general)
  • Depth discontinuities (and edges)
  • Narrow streets
  • Moving objects
  • Forest in leaf-off season
  • Reconstruction of small objects (wrt GSD)
Conclusions

- Provide **small test data dealing with these cases** (and keep them available!)
  - 1 strip with 5 images (80%)
  - or 3 strips (50%) with 5 images (80%)
  - undistorted images
  - (optional) provide a mask which focuses on the relevant problem zone(s)

- Software companies already offer evaluation and timely limited licences
  → Users (and companies!) can test existing software quickly with these open test data
Idea: Test data for forest (loosely grown)

OP with mask

DIM software 1

DIM software 2

ALS Ground Truth

Images adapted from: Ressl et al., 2016. Dense Image Matching vs. Airborne Laser Scanning – Comparison of two methods for deriving terrain models. PFG