VIENNA UNIVERSITY OF TECHNOLOGY
DEPARTMENT OF GEODESY AND GEOINFORMATION
RESEARCH GROUPS PHOTOGRAMMETRY & REMOTE SENSING

Riva del Garda - ISPRS
Technical Commission 2
Symposium - June 7, 2018

ASSESSING THE ACCURACY OF DENSE IMAGE MATCHING

(or Benchmarking DIM)

Camillo Ressl

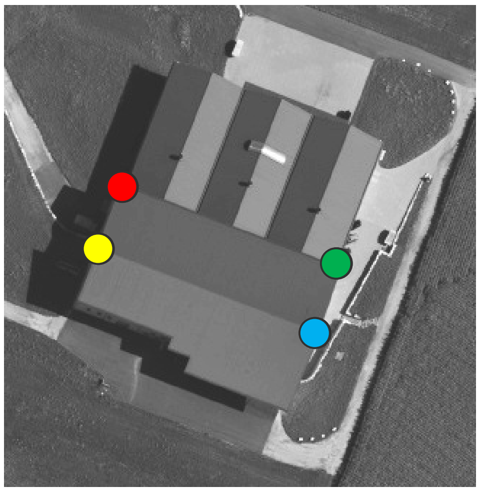
camillo.ressl@geo.tuwien.ac.at

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

Content

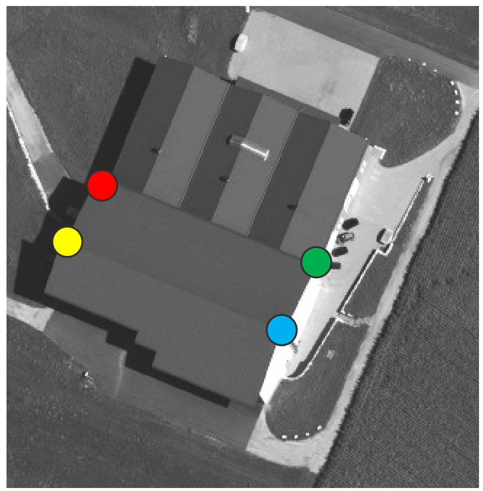
- Image Matching: Finding corresponding pixels in ≥ 2 images with given orientation

Image 1



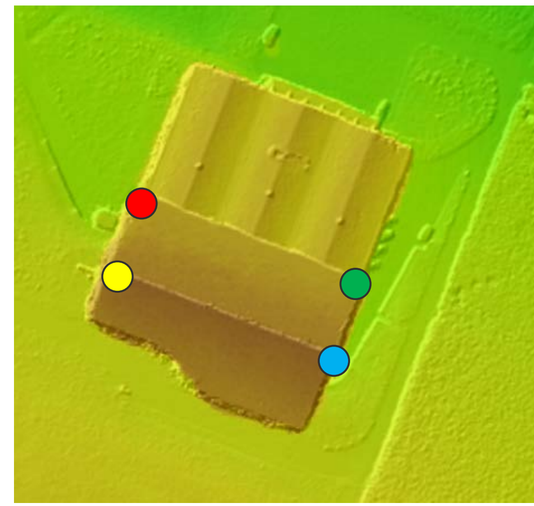
Vaihingen, DMC, 8cm, #71+73

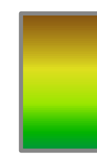
Image 2



spatial intersection

DSM





high

low

- Dense: match every pixel
- Result: point cloud (or 2.5D model) of object

Semi-Global Matching (SGM)

- In Computer Vision (CV): Hirschmüller, 2008

328 IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 30, NO. 2, FEBRUARY 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.

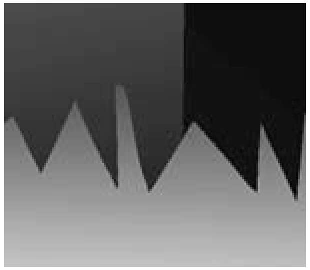
- In Photogrammetry:
  2011 first implementations appeared: Match-T, Sure

Benchmarks

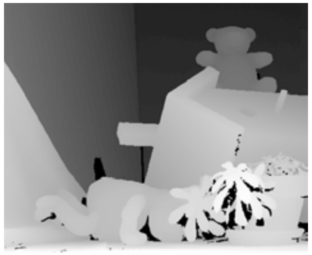
- Benchmarks are the processes and the results of assessing performance.
- For practice:
  - What accuracies can be achieved?
  - Which parameters influence the accuracy?
- For developers and researchers:
  - Which parameters (cost function, minimization method, …) perform best on various scenarios?
  - 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

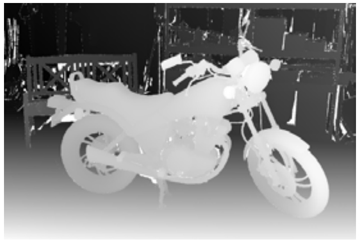
- http://vision.middlebury.edu/stereo/ → image pair
  training + evalutation images, upload result
- 2001: (0.2 MP)
  - scenes with planar objects
  - Ground truth (GT) disparties labled by hand

- 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

Scharstein & Szeliski, 2002, A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms, IJCV, 47.

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

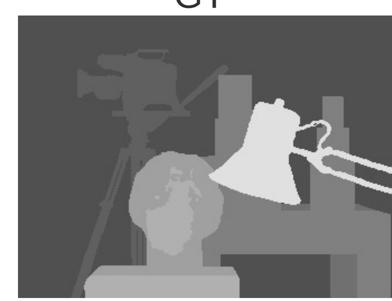
Middlebury (2 views)

- Quality measures:
  - RMS of disparity differences ΔD (per view)
  - Number of bad pixels: |ΔD| > 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)
- Participants: 16 (2002), 160 (2015)

Tsukuba scene (1996)

Test

GT

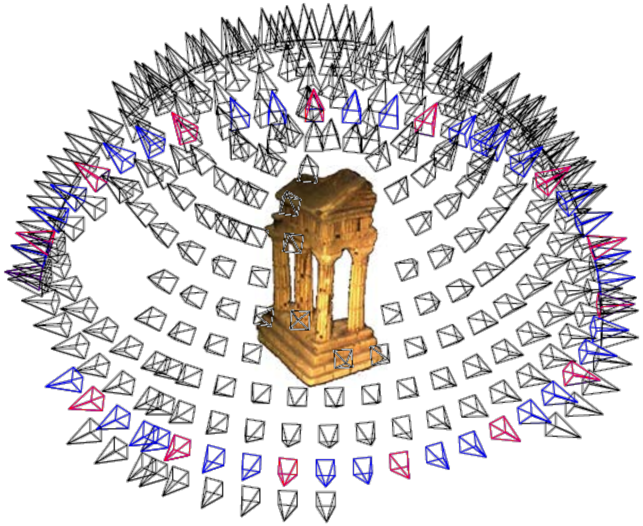


Scharstein & Szeliski, 2002, A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms, IJCV, 47.

Middlebury (multi view)

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

- http://vision.middlebury.edu/mview
- ~ 300 images by robot arm (GSD 0.25 mm)



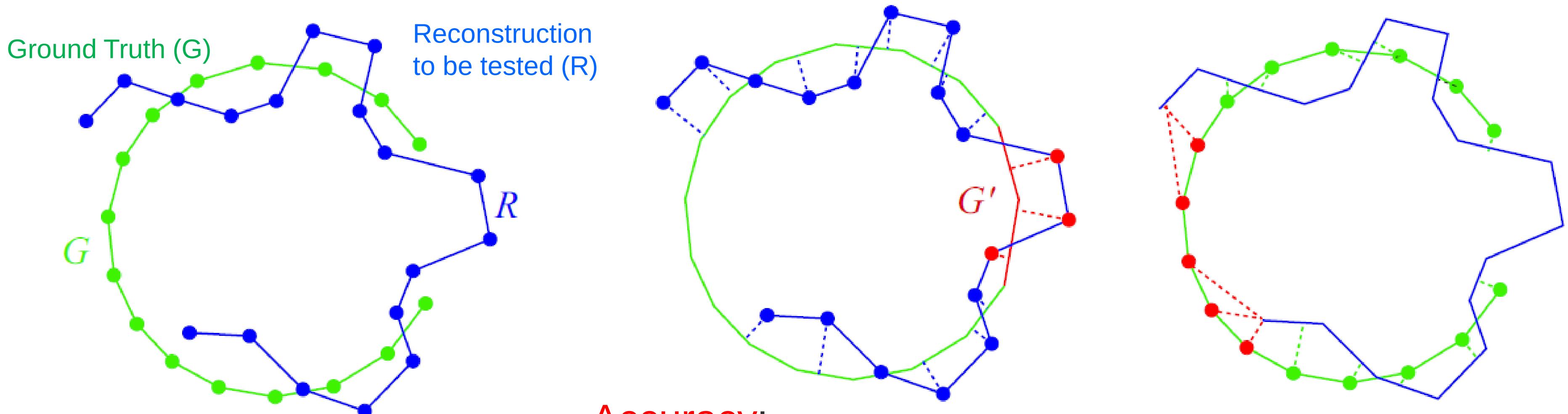
Ground Truth (GT)

- 4 different objects
- GT using laser strip scanner; alignment with images using ICP and maximizing photo-consistency
- Participants: 84 (2018)

Middlebury (multi view)

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

- 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



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."
  [Morales & Klette, 2010, Ground Truth Evaluation of Stereo Algorithms for Real World Applications, ACCV Workshops]
- "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]

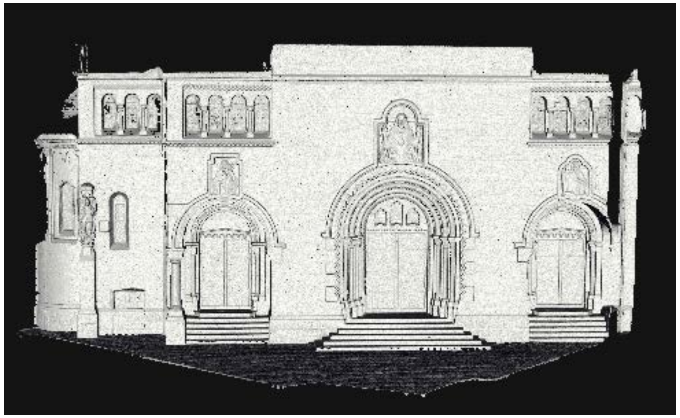
→ evaluate in the real-world

Multi-View Outdoor

Strecha, von Hansen, Van Gool, Fua, Thoennessen, 2008, On Benchmarking Camera Calibration and Multi-View Stereo for High Resolution Imagery, CVPR 08

- 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

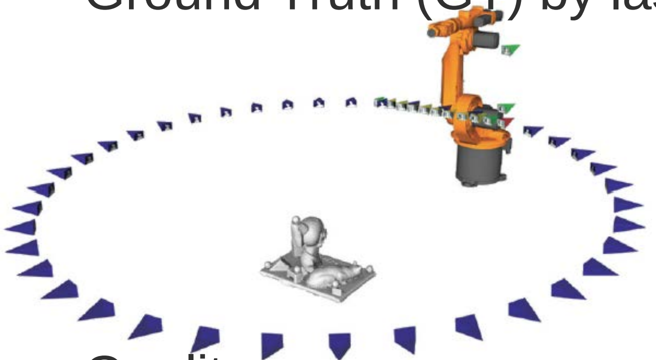


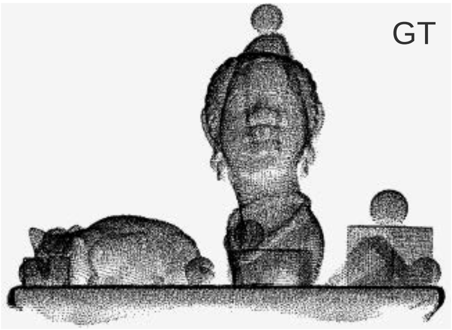
- Participants: 12 (2009), images still available

ISPRS WG III/2 2004 - 2008

- Data sets for shape-from-X
- X = stereo, motion, silhouette, shading, …
- Images using robot arm
- Ground Truth (GT) by laser scanner

Bellmann, Hellwich, Rodehorst, Yilmaz, 2007. A Benchmarking Dataset for Performance Evaluation of Automatic Surface Reconstruction Algorithms, CVPR '07





The benchmarking scene object

- Quality measures:
  - Accuracy = RMS of depth map differences ΔD (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?

DGPF 2009

- 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

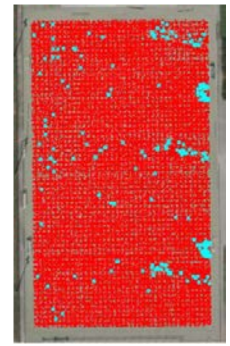
East-West ~ 8 km

DGPF 2009

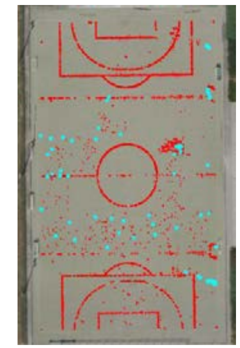
Haala, Hastedt, Wolf, Ressl, Baltrusch, 2010: Digital Photogrammetric Camera Evaluation – Generation of Digital Elevation Models, PFG 02/2010

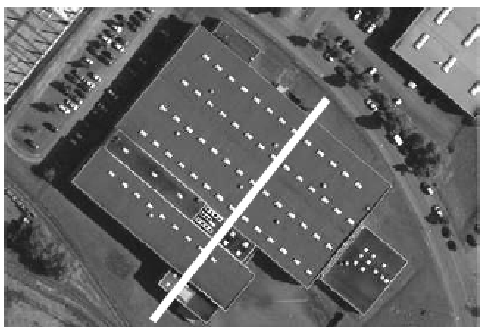
- Quality Measures:
  - GPS points vs. matched DEM, RMS = 0.4 - 1 GSD
  - soccer field: STD(robust) of matched points to common plane, STD = 0.3 - 2 GSD
  - profiles

DMC

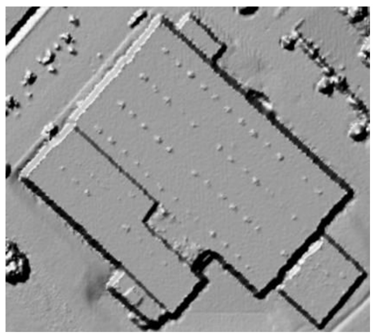


RMK

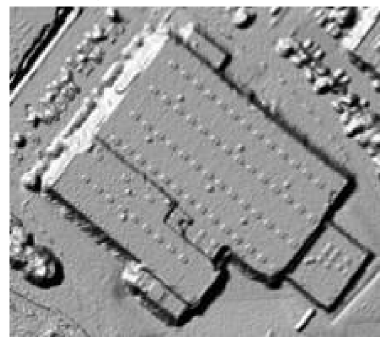




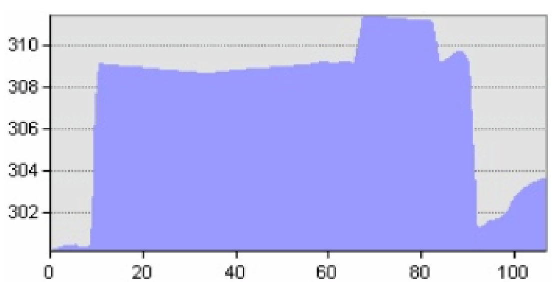
ALS



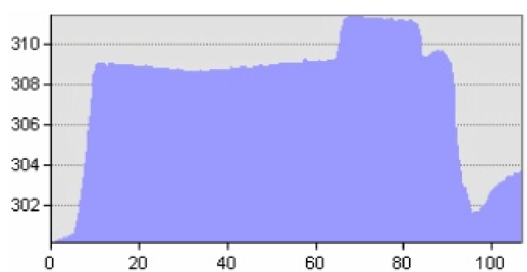
DMC, 8cm



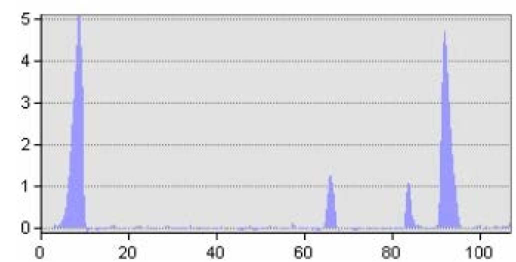
ALS



DMC

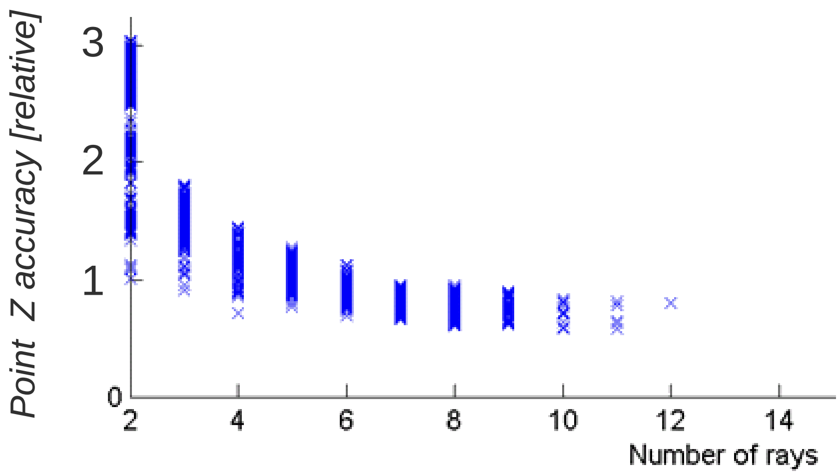


ALS - DMC



Our first DIM assessments (2011)

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



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

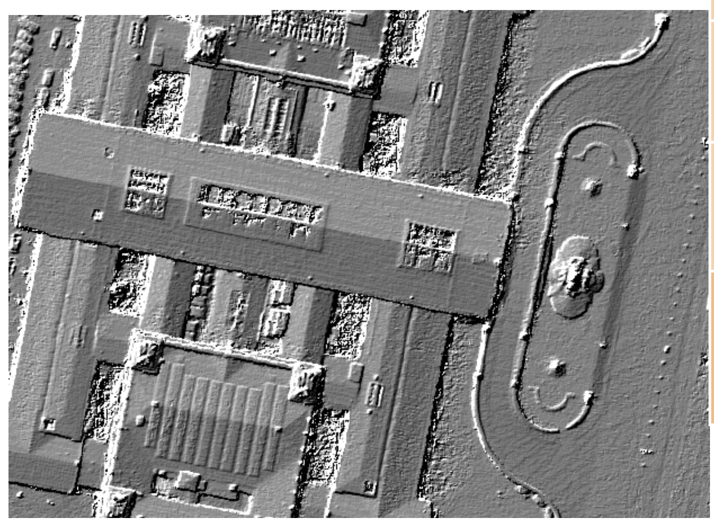
Height accuracy:
σ_{MAD} ~ 4.5 cm
(wrt ALS DSM)

Robust version of STD; see next slide.

pair(s)	n	σ_{MAD} [cm]
80%	1	**15**
60%	1	10
40%	1	8
80% Fusion	4	**8.4**
60% Fusion	3	7.1
80%+60% Fusion	7	6.6
80%+60% all strips + cross mods	<70	**4.5**

80% single pair

80% + 60 % fusion



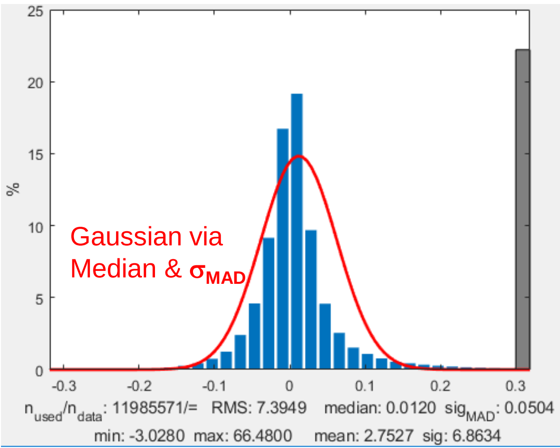
In cooperation with Stadt Wien, MA41.

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?

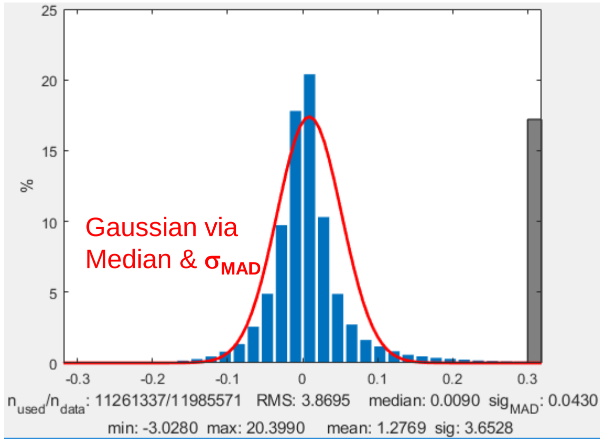
all dz values



Mean = 2.8 m
σ = 6.9 m
Median = 1.2 cm
σ_{MAD} = 5.0 cm

σ_{MAD} = 1.4826 MAD
MAD(x) := median of absolute distances to median(x)

|dz| < 3 * σ



Mean = 1.3 m
σ = 3.7 m
Median = 0.9 cm
σ_{MAD} = 4.3 cm

→ σ_{MAD} only applies if distribution is Gaussian

→ Do not use the 3 * sigma rule, if you do not know sigma (and expectation) !

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

Chebyshev's inequality:

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

k = 3 → P = 11%

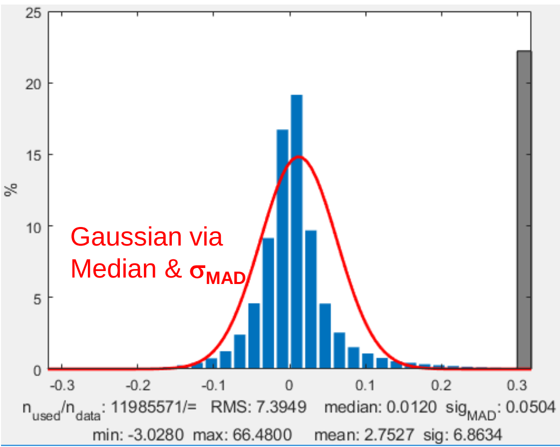
e.g. if you have 100 values with 21 outliers and you apply the 3*sigma rule (with sigma estimated from this corrupted sample), then afterwards you will have still at least 10 of these outliers.

nota bene: Use robust statistics, … but take care !

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

What is accuracy of last echos?

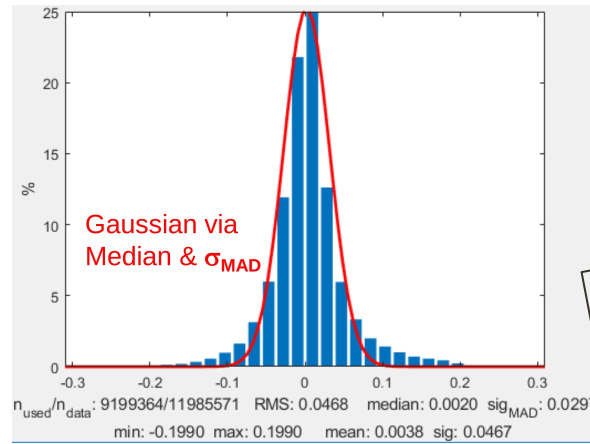
all dz values



Mean = 2.8 m
σ = 6.9 m
Median = 1.2 cm
σ_{MAD} = 5.0 cm

σ_{MAD} = 1.4826 MAD
MAD(x) := median of absolute distances to median(x)

|dz| < 4 * σ_{MAD}



Mean = 0.4 cm
σ = 4.7 cm
Median = 0.2 cm
σ_{MAD} = 3.0 cm

Histogram still not Gaussian!

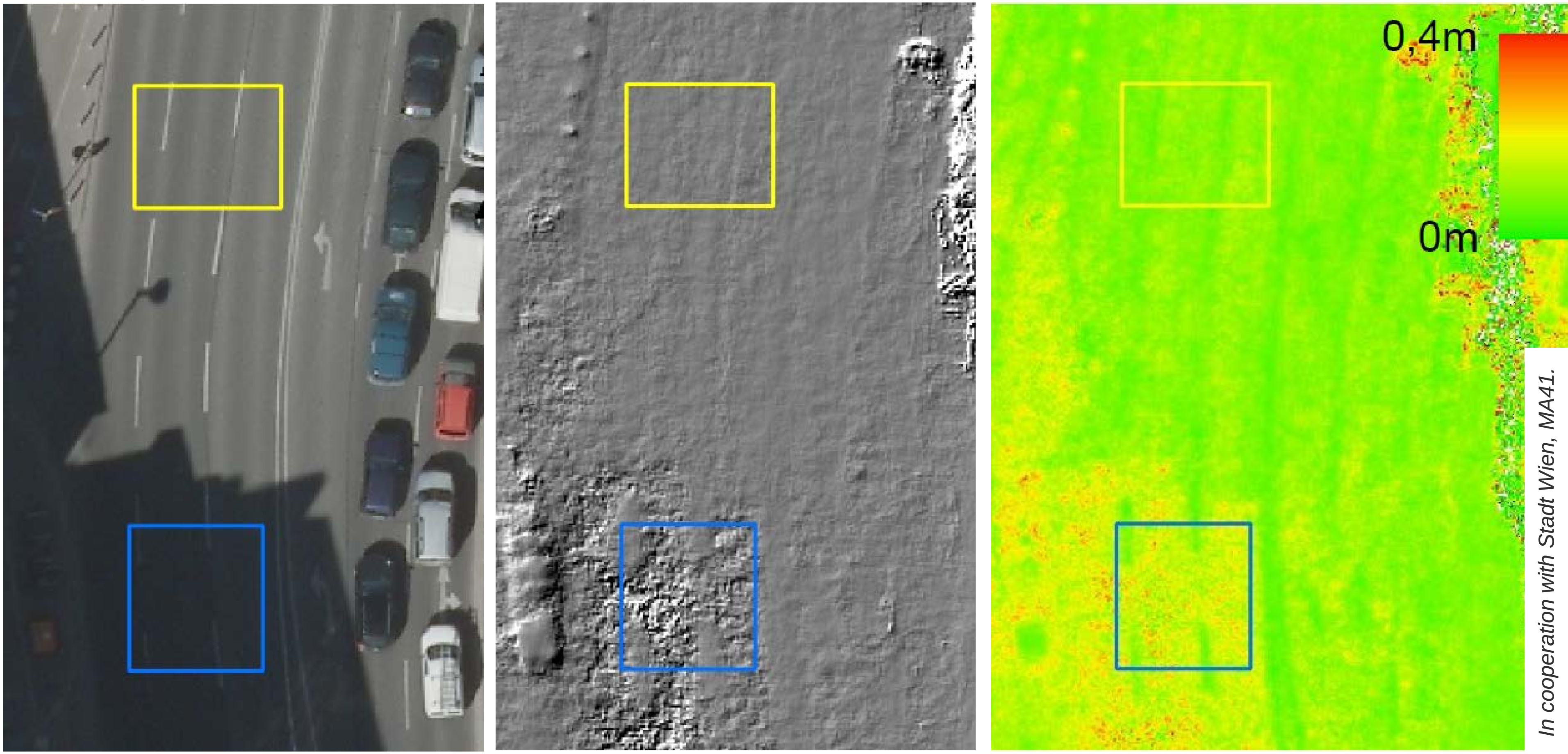
→ σ_{MAD} only applies if distribution is Gaussian

Take limits from the histogram directly, or apply k* σ_{MAD} (with k = 3 or 4 to start with).

In publications, please, report Mean, σ, and Median, σ_{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(max-min): 21cm (sun) vs. 87cm (shadow)

dZ(std): 4cm (sun) vs. 11cm (shadow)

→ should smoothing/regularization be based on scene content ?

Problems at homogenous texture and corners

Shading (MatchT, SGM, fusion)

aerial image

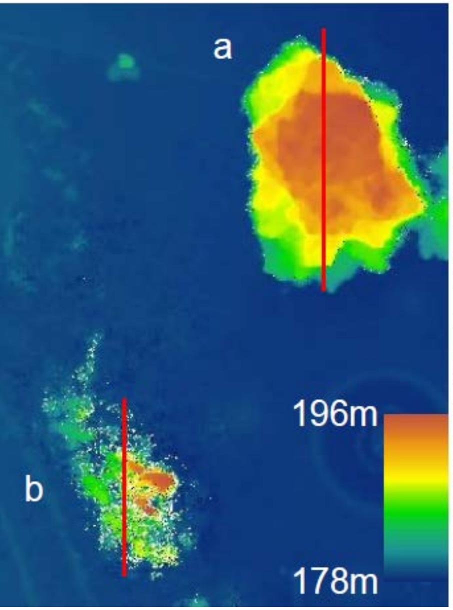
In cooperation with Stadt Wien, MA41.

Matching of trees

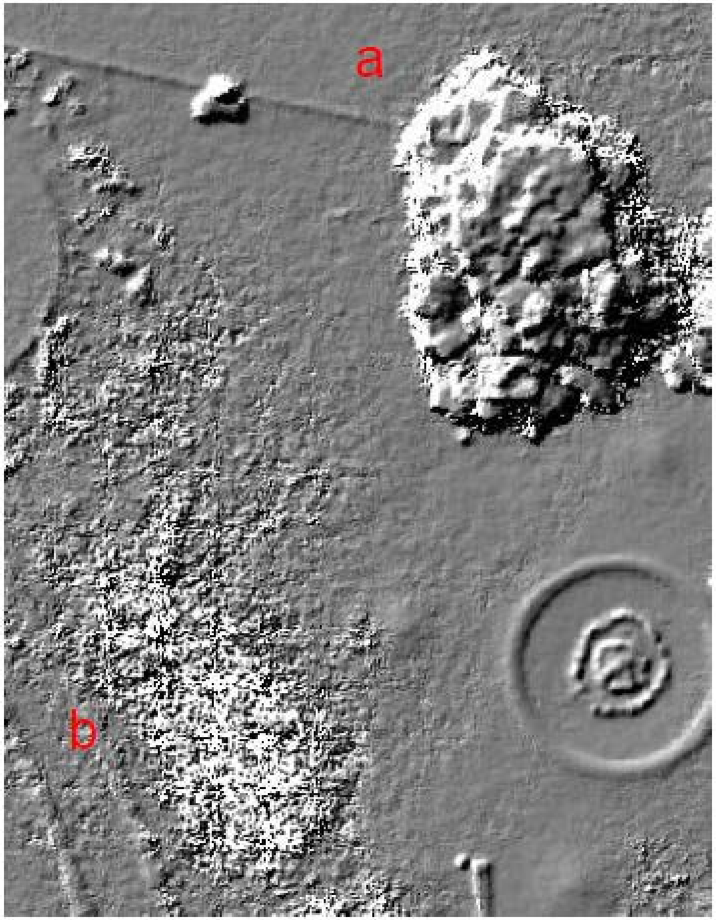
a: conifer

b: deciduous tree (leafless)

Z-coding



shading (MatchT, SGM, fusion)

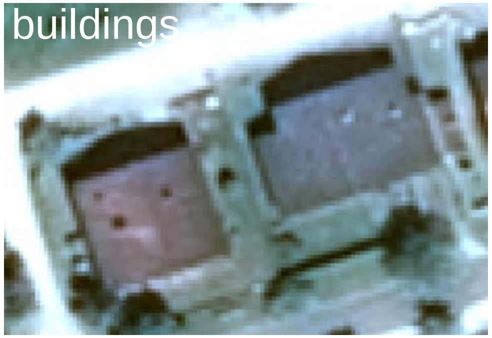


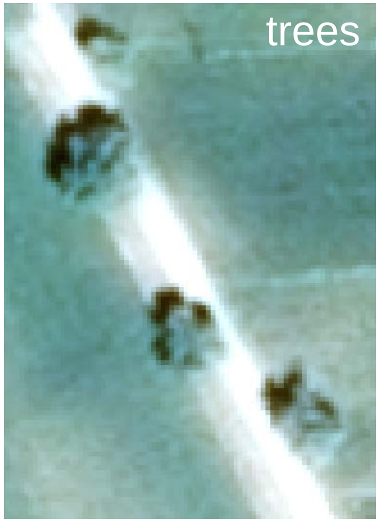
aerial image

In cooperation with Stadt Wien, MA41.

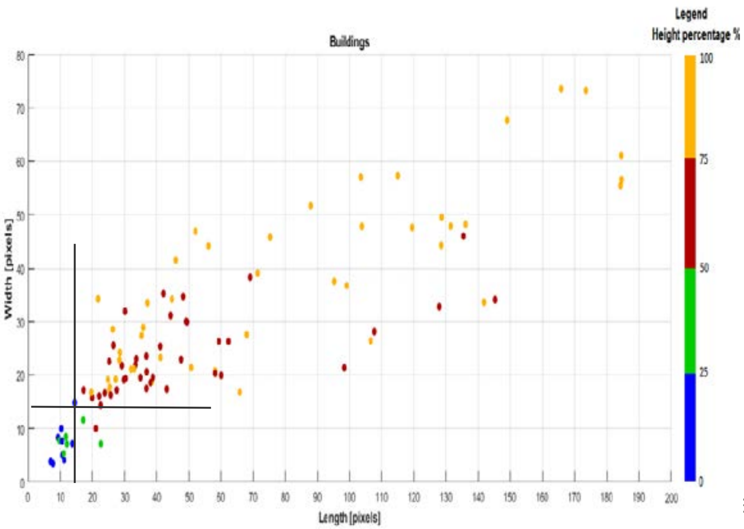
Reconstruction of isolated objects from Satellite

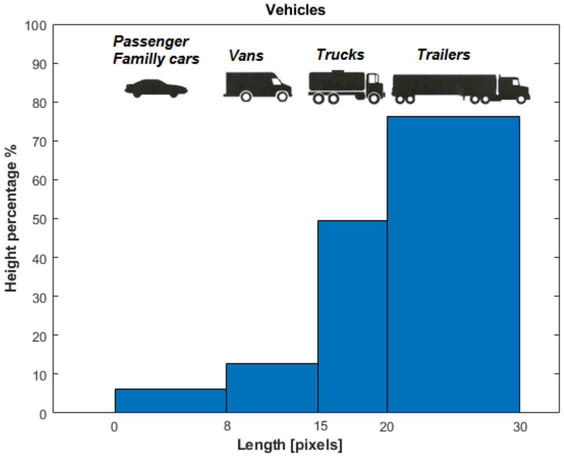
- Only 2-3 images (Pleijades, GSD 0.5 m)
- Reconstruct by matching these objects (Match-T):

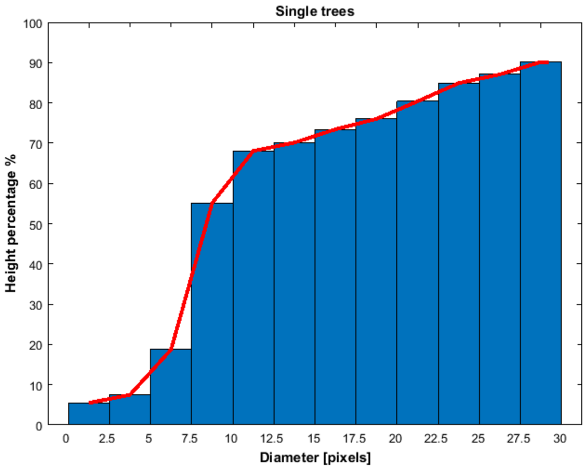




Height comparison with manual reconstruction: object size ≥ 15 pix → Z correct >50%







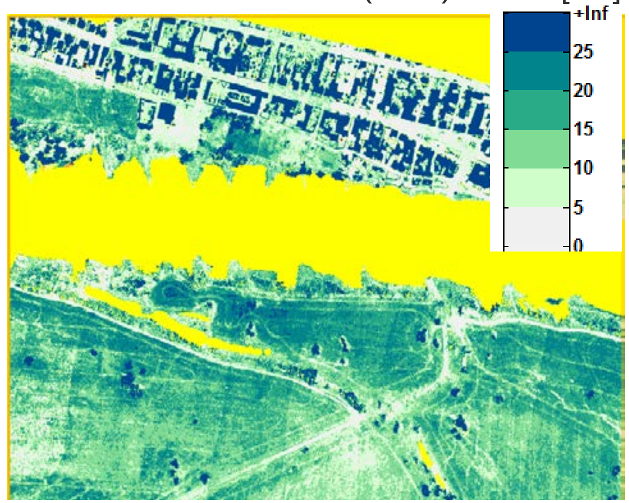
Effect of Land Cover on Height Accuracy

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

Ressl, Brockmann, Mandlburger, Pfeifer, 2016. Dense Image Matching vs. Airborne Laser Scanning – Comparision of two methods for deriving terrain models. PFG

OP

DIM – ALS(DTM)



3 land cover classes

	Sealed		Vegetation		Vegetation	
[cm]	med	std	med	std	med	std
ALS (last echos)	2.6	**5.5**	1.1	3.9	2.1	4.3
DIM	5.4	**3.7**	**19.7**	4.4	**17.7**	5.5

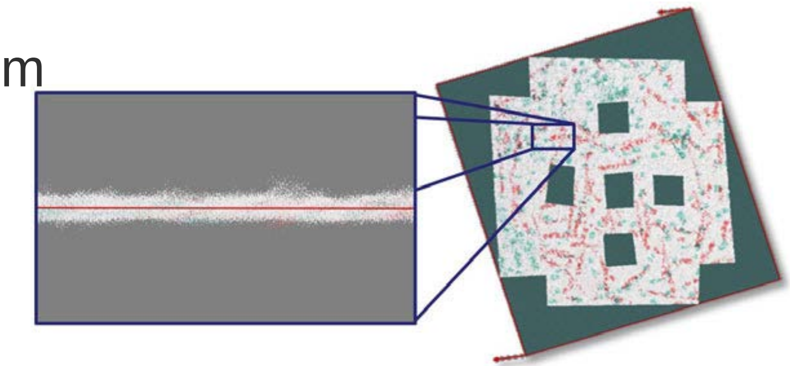
→ DIM over sealed areas better than ALS (bad SNR)

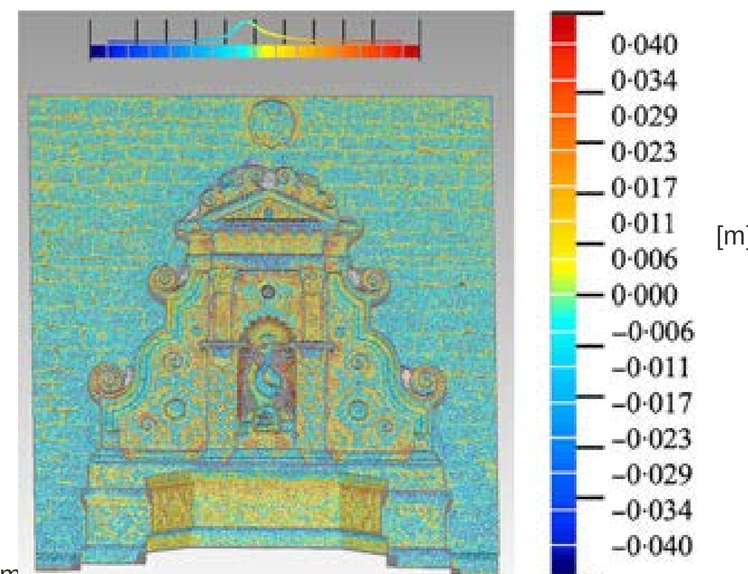
→ DIM at top of grass, ALS penetrates

Remondino et al., 2014

Remondino, Spera, Nocerino, Menna, Nex, 2014. State Of The Art In High Density Image Matching, The Photogrammetric Record

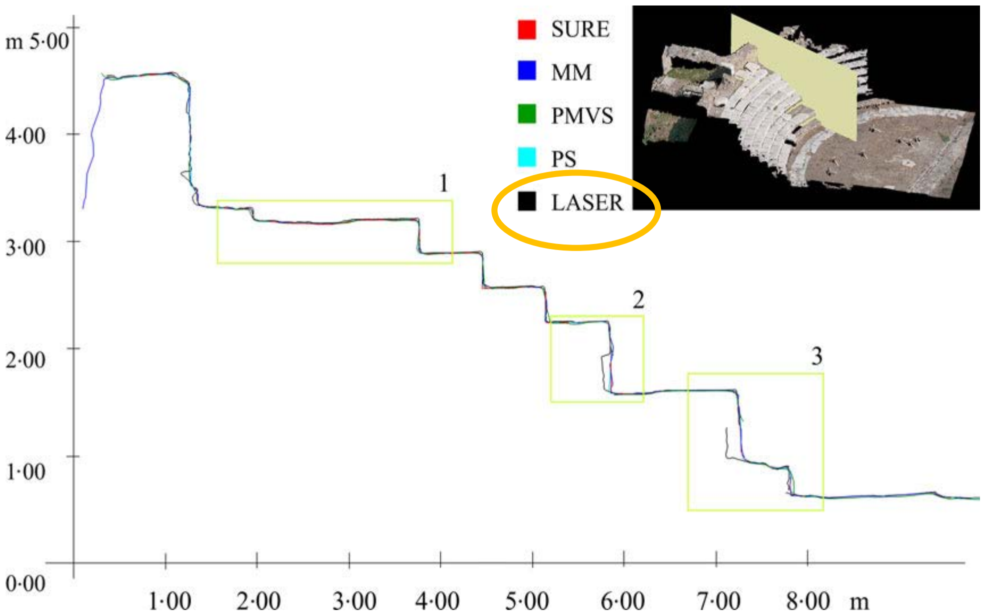
- Software tested:
  Sure, MicMac, PMVS, Photoscan
- 8 different objects: GSD = 0.06 mm to 12 cm
- Evaluation of point clouds (not meshes)
- Ground Truth: TLS
- Quality Measures:
  - Flatness: STD = 0.5 - 1 GSD
  - Comparison with TLS-mesh: STD = 0.5 - 1 GSD
  - Profiles
- Problems spotted:
  - Shadows
  - Small structures
  - Sharp discontinuities

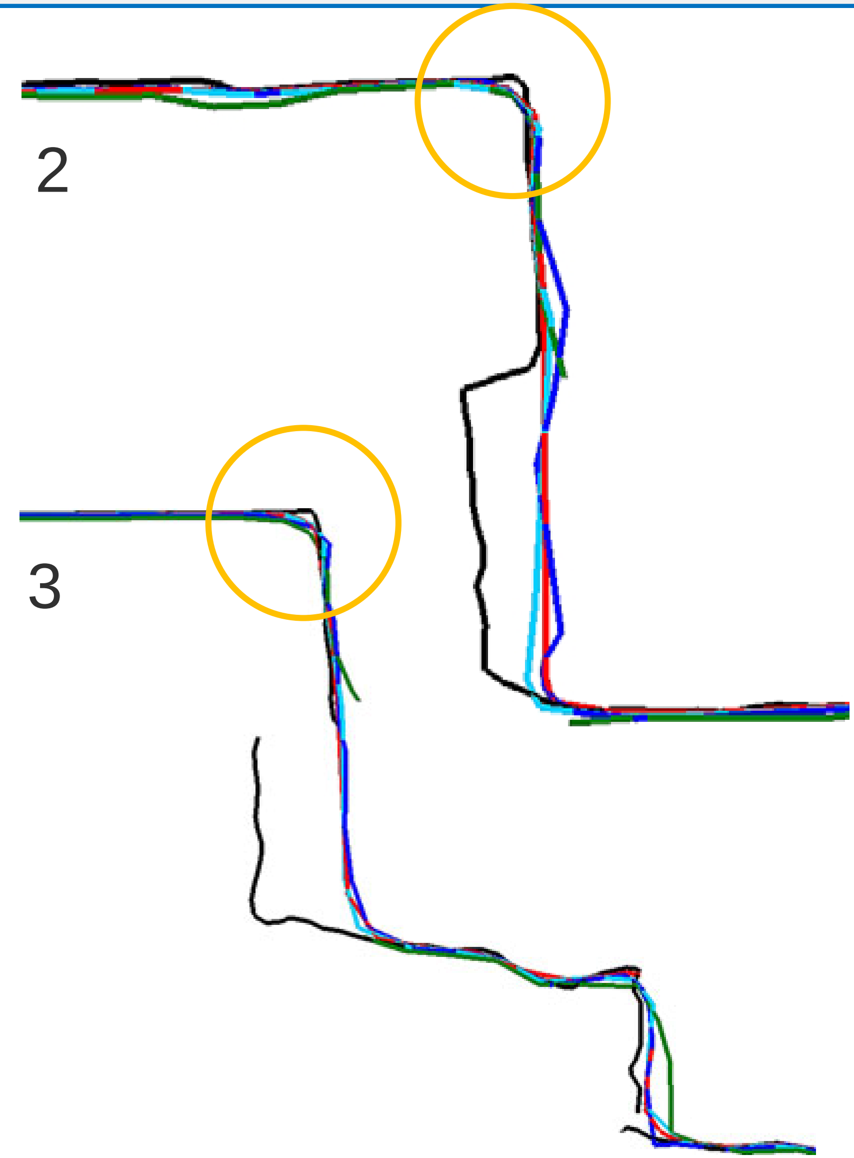




Remondino et al., 2014

- Problems at discontinuities



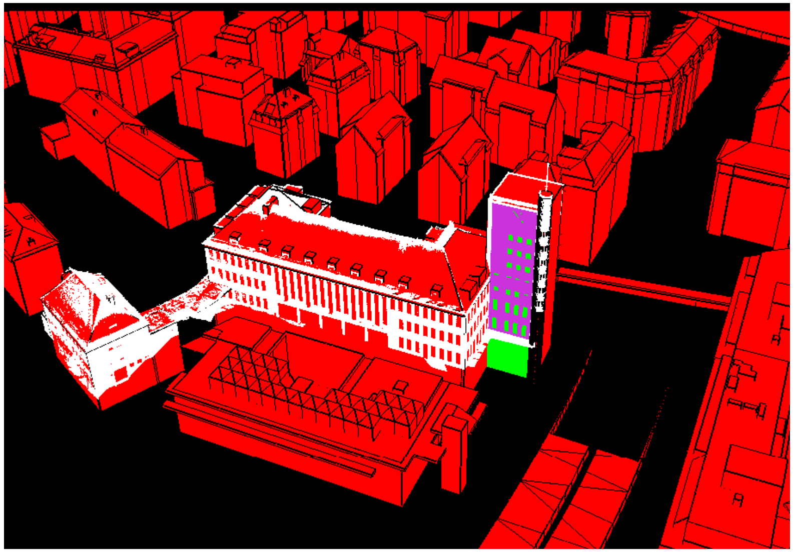


Remondino, Spera, Nocerino, Menna, Nex, 2014. State Of The Art In High Density Image Matching, The Photogrammetric Record

EuroSDR & ISPRS WG III/1 2012 - 2016

Cavegn, Haala, Nebiker, Rothermel, Tutzauer, 2014. Benchmarking High Density Image Matching For Oblique Airborne Imagery, ISPRS Archives

- **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



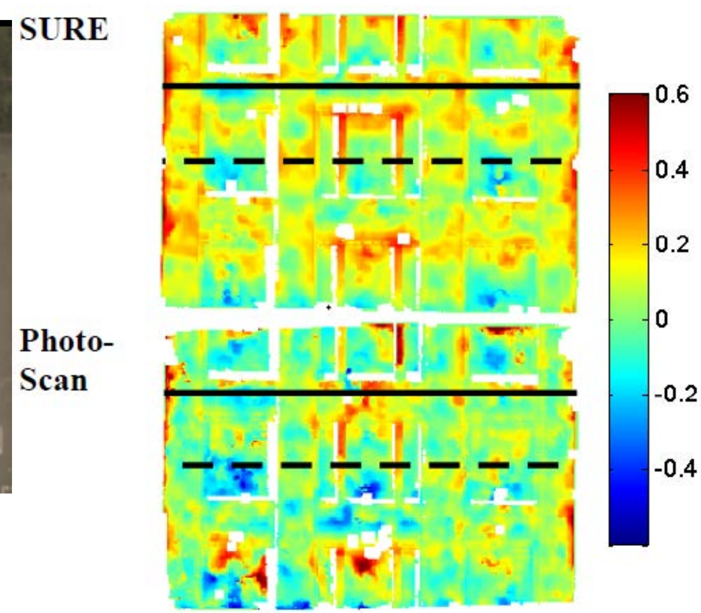
white: TLS point cloud, green: selected facade, purple: selected TLS point cloud

- Data still available: http://www.ifp.uni-stuttgart.de/ISPRS-EuroSDR/ImageMatching
- More data at: http://www2.isprs.org/commissions/comm1/icwg15b/benchmark_main.html

EuroSDR & ISPRS WG III/1 2012 - 2016

Cavegn, Haala, Nebiker, Rothermel, Tutzauer, 2014. Benchmarking High Density Image Matching For Oblique Airborne Imagery, ISPRS Archives

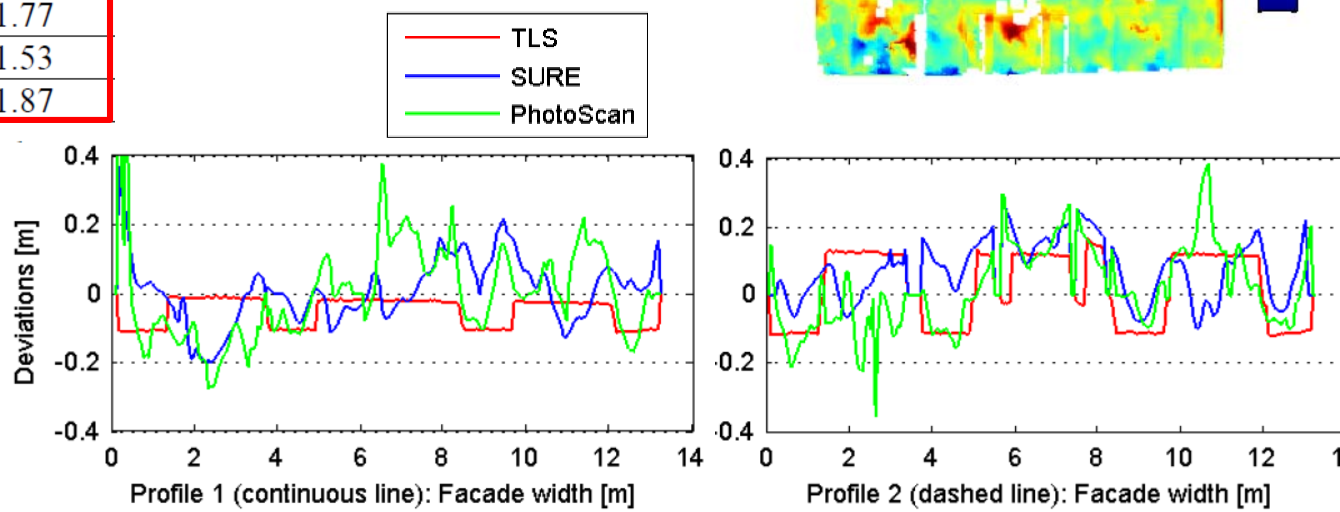
- Quality Measures:
  - Density
  - RMS of flatness error
  - DIM vs. TLS
  - Profiles



	GSD [cm]	Density [Points / m^2]	RMSE DIM [px]	RMSE DIM-TLS [px]
C S	7.7	120	1.56	1.92
C PS	7.7	172	2.06	1.77
R S	7.8	120	1.61	1.53
R PS	7.8	163	2.12	1.87

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 Photogrammetric community are a bit short-lived

  → petition: keep data available

- Computer Vision shifts towards real word scenes,
  maybe Photogrammetry should look for controlled environments ?

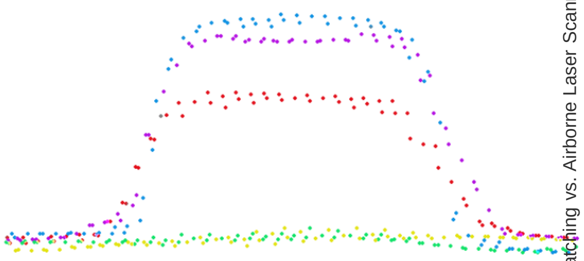
- Depth accuracy is well handled for nadir (oblique still open research)

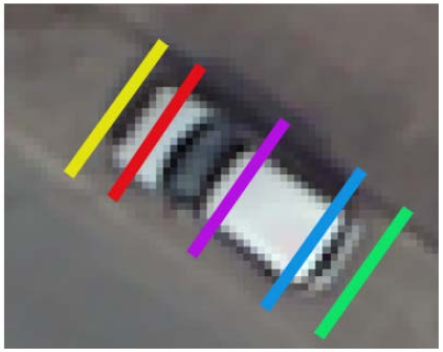
- Future:

  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)





Ressl et al., 2016. Dense Image Matching vs. Airborne Laser Scanning – Comparision of two methods for deriving terrain models. PFG

Conclusions

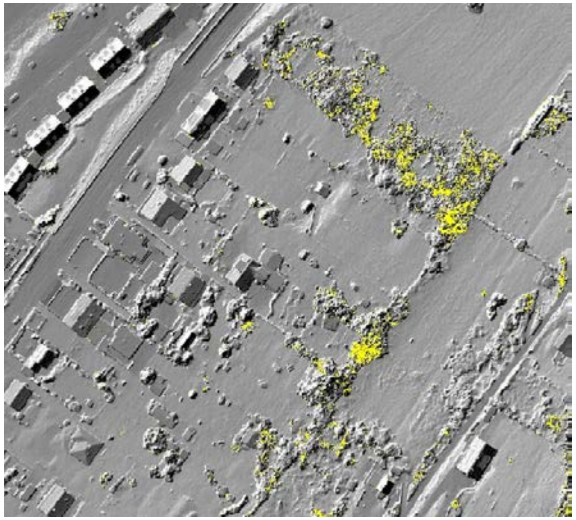
- Provide small test data dealing with these cases (and keep them available !)
  - 1 strip with 5 images (80%)
  - or 3 strips (50%) with 5 images (80%)
  - undistorted images
  - (optional) provide a mask which focuses on the relevant problem zone(s)

- Software companies already offer evaluation and timely limited licences
  → Users (and companies!) can test existing software quickly with these open test data

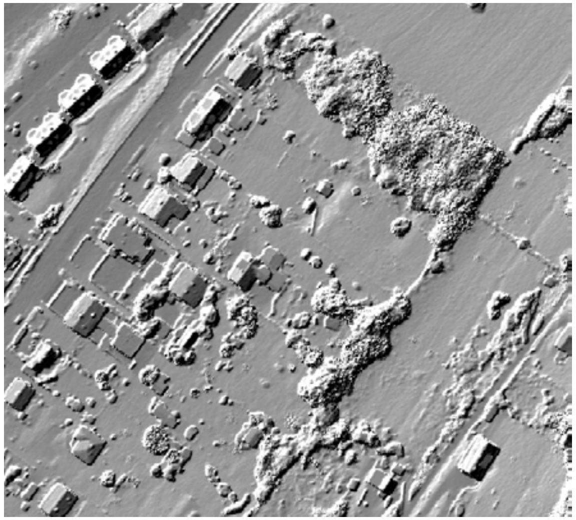
Idea: Test data for forest (loosely grown)

OP with mask

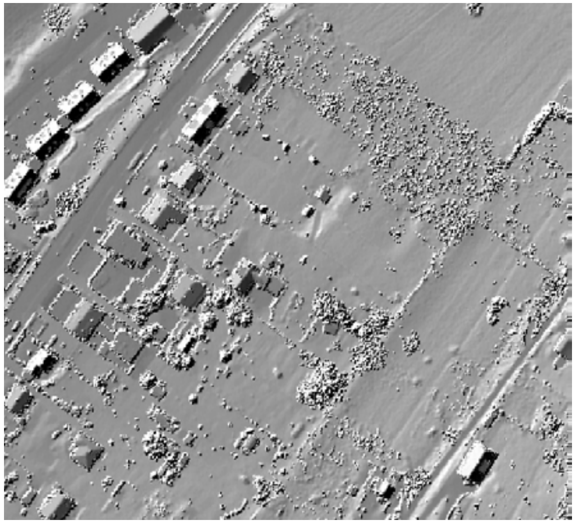
DIM software 1



DIM sotware 2



ALS Ground Truth



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