EFFICIENT AND GLOBALLY OPTIMAL MULTI VIEW DENSE MATCHING FOR AERIAL IMAGES

Arnold Irschara[a], Markus Rumpler[b], Philipp Meixner[b], Thomas Pock[b] and Horst Bischof[b]

[a] Microsoft Photogrammety, Anzengrubergasse 8, A-8010 Graz, Austria - Arnold.Irschara@microsoft.com
[b] Institute for Computer Graphics and Vision, Graz University of Technology, Inffeldgasse 16, A-8010 Graz, Austria - {rumpler,meixner,pock,bischof}@icg.tugraz.at

Commission III/4

KEY WORDS: Mapping, Vision, Processing, Aerial, High resolution

ABSTRACT:

A variety of applications exist for aerial 3D reconstruction, ranging from the production of digital surface models (DSMs) and digital terrain models (DTMs) to the creation of true orthophoto and full 3D models of urban scenes that can be visualized through the web. In this paper we present an automated end-to-end workflow to create digital surface models from large scale and highly overlapping aerial images. The core component of our approach is a multi-view dense matching algorithm that fully exploits the redundancy of the data. This is in contrast to traditional two-view based stereo matching approaches in aerial photogrammetry. In particular, our solution to dense depth estimation is based on a multi-view plane sweep approach with discontinuity preserving global optimization. We provide a fully automatic framework for aerial triangulation, image overlap estimation and dense depth matching. Our algorithms are designed to run on current graphics processing units (GPUs) which makes large scale processing feasible at low cost. We present dense matching results from a large aerial survey comprising 3000 aerial images of Graz and give a detailed performance analysis in terms of accuracy and processing time.

1 INTRODUCTION

Novel digital aerial cameras capture high resolution images that are readily suitable for the creation of photogrammetric end products like digital surface models (DSMs), digital terrain models (DTMs), orthophotos and full 3D models of urban scenes that can be visualized trough the web (Zebedin, 2010). Fully automated image based creation of dense point clouds with an elevation measurement at each pixel is nowadays feasible at low cost and makes the technology competitive with LiDAR-based surface measurements (Leberl et al., 2010). It can be argued that the image-based approach offers many advantages over LiDAR, and that practically all aerial mapping scenarios will need digital images, even with LiDAR (Leberl and Gruber, 2003). In the last two decades digital aerial cameras reached a mature technical state and hence provided the necessary geometric and radiometric stability and resolution to compete with analogue aerial cameras. Furthermore, they feature the advantage that the acquired digital images are readily suitable for automatic processing. The high redundancy of multiple overlapping images holds the promise of full automation of the matching process. Therefore, many essential products obtained from aerial images can be cost-efficiently derived by an automated processing pipeline. Today, airborne photogrammetric surveys are flown with at least 80% forward overlap and 60% side-lap. The high resolution imagery opens the possibility to generate detailed maps of an environment. Figure 1 depicts a typical reconstruction result of a building obtained from aerial images. A standard aerial mapping project consists of 500 to 20,000 high resolution images, which translates into 0.3-12 terabyte of raw data. Under the premise that processing of such a massive amount of data can be done in reasonable time, passive photogrammetry can directly outperform current LiDAR systems (Baltsavias, 1999) by means of ground sampling distance (GSD) and reduced flight costs (Leberl et al., 2010). This directly leads to significant economic benefits. With a further increase of ground sampling distance to the range of about 1cm, aerial photogrammetry could even compete with traditional ground measurement devices such as surveys based on total stations and GNSS/GPS systems.

Figure 1: Visualization of a textured 3D reconstruction/DSM derived from our multi-view dense matching method.

1.1 Aerial Images

Aerial images produced by state of the art large format camera systems such as the products from Vexcel Imaging[1] currently comprise up to 260 megapixels (UltraCam Eagle) at a high radiometric resolution (Figure 2). Aerial flight missions are normally performed using GPS-Aided Inertial Navigation that allows Direct Georeferencing (Hutton and Mostafa, 2005), hence Aerial Triangulation (AT) and ground control points are not necessarily required. However, GPS-Aided Inertial Navigation has several requirements that makes such systems costly and hard to apply in real world. First of all, the IMU must be rigidly attached to the camera and any misalignment of IMU/camera needs to be calibrated. Second, exact time of image exposure and GPS/INS pose must be provided. Third, the camera interior geometry (focal length, principal point) must be well calibrated and stable. Even if calibration is done accurately, total reliance on GPS/IMU

[1] http://www.microsoft.com/ultracam

does compromise the accuracy of resulting stereo matches and point cloud patches (Leberl et al., 2010). In this paper we employ a fully automated processing pipeline that computes the scene structure and camera orientations from aerial input images. In computer vision, this approach is known as Structure from Motion (SfM) and delivers subpixel accurate photo alignment from image measurements, only. Hence this method is more flexible than current photogrammetric systems that require GPS/IMU and the semi-automatic selection of point measurements in overlapping images.

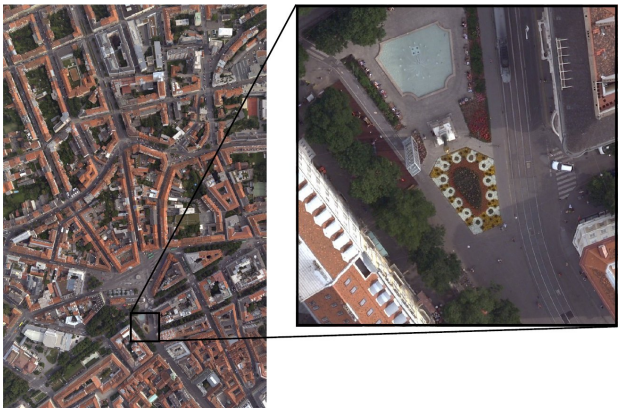


Figure 2: High resolution aerial image comprising 7500×11500 pixels at a ground sampling distance (GSD) of 10cm/pixel.

1.2 Aerial Triangulation

We employ a fully automated processing pipeline that computes the scene structure and camera orientations from aerial input images, only. First, several thousand Points of Interest (POIs) are extracted from each image using the Scale Invariant Feature Transform (SIFT) (Lowe, 2004). Next, features between pairs of adjacent images along the flight path are matched. Given the image sequence $\mathcal{I}$ with n images, $\mathcal{I} = \{I_t | t = 1, \ldots, n\}$ the features of each view I_t are matched with a number of adjacent views I_{t+i} with $i = \{-r, \ldots, +r\}$ and $i \neq 0$ where r determines the matching interval. We use $r = 5$ to match aerial images with a forward overlap of 80%. This method achieves tracks along the flight paths but might miss correspondences between flight lines. To establish correspondences between flight lines, an image retrieval approach based on a vocabulary tree search (Nistér and Stewenius, 2006) is performed. Such an approach assumes that each image is represented as a bag of words (Sivic and Zisserman, 2003) and the employed method efficiently determines a similarity score of all image pairs. In general, overlapping images achieve a higher score than unrelated images, hence this approach is able to detected potential matching candidates across flight lines. We use exhaustive SIFT descriptor matching between pairs of frames and use the Lowe distance ratio criterion (Lowe, 2004) to achieve matching uniqueness. Next, the Five Point relative pose algorithm (Nistér, 2004) inside a RANSAC loop (Fischler and Bolles, 1981) is used to robustly compute pairwise camera orientations. The output of the automatic matching procedure is a graph structure denoted as epipolar graph $\mathcal{EG}$, that consists of the set of vertices $\mathcal{V} = \{V_1 \ldots V_N\}$ corresponding to the images and a set of edges $\mathcal{E} = \{e_{ij} | i, j \in \mathcal{V}\}$ that are pairwise reconstructions, i.e. relative orientations between view i and j, $e_{ij} =< P_i, P_j >$,

$$P_i = K_i[I, 0] \text{ and } P_j = K_j[R, t], \tag{1}$$

where P_i, P_j are respective projection matrices. The epipolar graph $\mathcal{E}G$ encodes relative orientations and pairwise reconstructions. Chaining all relative orientations together should result in a global consistent 3D structure. We follow a greedy, incremental reconstruction approach (Snavely et al., 2008) to iteratively reconstruct the scene from an initial image pair. Structure and camera pose refinement is done using robust bundle adjustment (Triggs et al., 2000). Figure 3 illustrates an orientation result of 3000 aerial images reconstructed with our fully automated aerial triangulation framework.

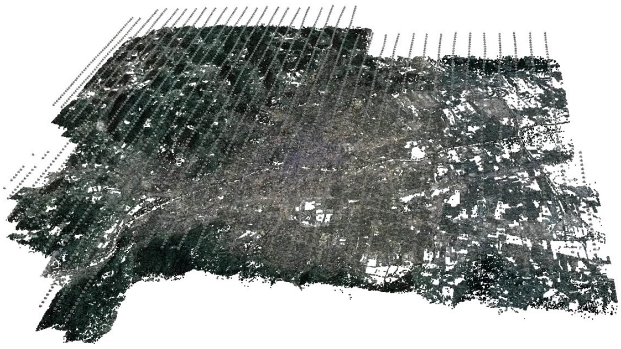


Figure 3: Aerial Triangulation (AT) result from 3000 aerial images covering an area of approximately 150km^2 of Graz and surrounding.

1.3 Multi-View Reconstruction

While traditional dense matching approaches for aerial images are based on stereo pairs (Hirschmüller, 2006), we employ a multi-view method for dense depth map extraction. The high redundancy of image overlap assists in the correspondence problem and allows to overcome some of the shortcomings of traditional stereo. In contrast to stereo, multi-view scene reconstruction provides additional information by capturing a scene from different viewpoints. From a geometric point of view, a triangulation angle of 90° delivers the highest accuracy, hence wide baseline settings are preferred. The depth uncertainty for a canonical stereo pair can be directly derived from the disparity error,

$$\epsilon_z = \frac{bf}{d} - \frac{bf}{d + \epsilon_d} \approx \frac{z^2}{bf} \epsilon_d, \tag{2}$$

where z is the point depth, f the focal length and b the image baseline. This means that depth precision is mainly a function of the ray intersection angle (Gallup et al., 2008) α, having a minimum at $\alpha = 90^\circ$. On the other hand, a large parallax introduces occlusion and perspective distortions which makes matching a challenging problem.

Given a photogrammetric network of cameras and correspondences with known error distribution, the precision of triangulated points can be determined from the 3D confidence ellipsoid (i.e. covariance matrix), as shown in (Beder and Steffen, 2006). Figure 4 shows an evaluation of the uncertainty ellipsoid of triangulated points for the UltraCamD at flying height $900m$ with 80% forward overlap and 60% sidelap. A multi-view triangulation from 20 image measurements achieves an accuracy of about $\sigma_z = 5$cm in depth and about $\sigma_{x,y} = 1.8$cm for in-plane measurements. As shown in (Rumpler et al., 2011) true multi-view matching outperforms two view stereo approaches by about one order of magnitude in terms of achievable geometric accuracy. Overall, multiple views for 3D reconstruction contribute to the scene completeness and increase scene coverage by capturing areas that might be occluded in traditional stereo.

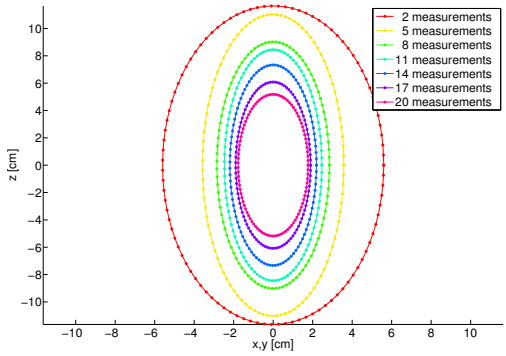


Figure 4: Covariance ellipsoid for one 3D point depending on the number of image measurements used for multi-view triangulation.

2 DENSE MATCHING

Structure from motion yields camera orientations and a sparse set of triangulated points, but for photogrammetric end-products like ortho-image creation or Digital Surface Model (DSM) extraction, dense 3D geometry is required. Our solution to dense depth estimation is based on a multi-view plane sweep approach (Collins, 1996) with global optimization on a 3D voxel space.

2.1 Multi-View Plane Sweep

Plane sweep techniques in computer vision are simple and elegant approaches for image based reconstruction with multiple views, since image rectification is not required. The 3D space is iteratively traversed by parallel planes which are usually aligned with a particular key view (Figure 5).

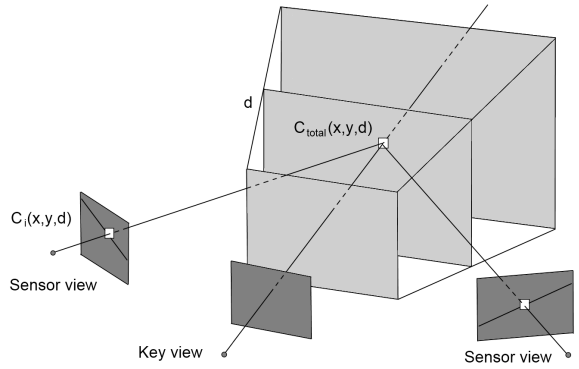


Figure 5: The scene is traversed by planes parallel to the key view. For each discrete depth, sensor images are projected onto the plane and similarity is compared between pixels of key and sensor views by calculating cost values $C_i(x, y, d)$. A cost volume is filled with the accumulated matching scores.

The plane at a certain depth d from the key view induces homographies for all other views, thus the sensor images are warped to the current plane $\pi = (n^\top, d)$. Here, n is the plane normal and d the current depth hypothesis. The key view is assumed in canonical coordinates $P = K[I \mid 0]$ according to the appropriate homography,

$$H = K' \left(R - \frac{tn^\top}{d} \right) K^{-1}, \tag{3}$$

that transfers coordinates x from the sensor view to image positions x' of the key view with $x' = Hx$. Here, K is the intrinsic matrix of the key view and R, t is the relative pose of the sensor view $P' = K'[R \mid t]$ with respect to the key view. Given two projection matrices $P_1 = K_1[R_1 \mid t_1]$ and $P_2 = K_2[R_2 \mid t_2]$ the relative pose between P_1 and P_2 is computed from,

$$R = R_2 R_1^\top, \tag{4}$$

$$t = t_2 - R_2 R_1^\top t_1 \tag{5}$$

and the normal vector of the plane $n = [0, 0, -1]$. If the plane at a certain depth passes exactly through the surface of the object, the color values from the key view and from the mapped sensor views should coincide at appropriate positions. By sweeping the plane through the 3D space, a cost volume is filled with image correlation values that corresponds to the disparity space image (DSI) in traditional stereo (Seitz et al., 2006).

2.2 Initialization

Image-space algorithms usually constrain the maximum disparity range or interval, in which depth values can occur. Respectively, the extent of scene geometry is determined to lie between a near and far plane from the camera center of a key view (Figure 6). Minimal and maximal depth range $[z_{near}, z_{far}]$ can either be estimated from the sparse scene reconstruction from SfM or explicitly set to a global value if prior knowledge about the minimum and maximum scene depth is available, e.g. from a coarse digital surface model. Such a model may be already available through previous aerial mapping surveys, or alternatively, can be generated by combining multiple public domain geographic information sources.

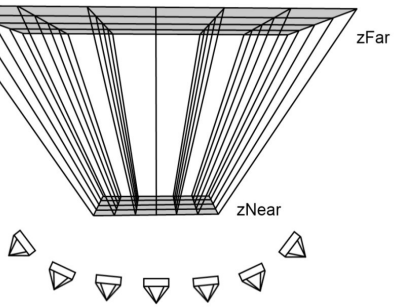
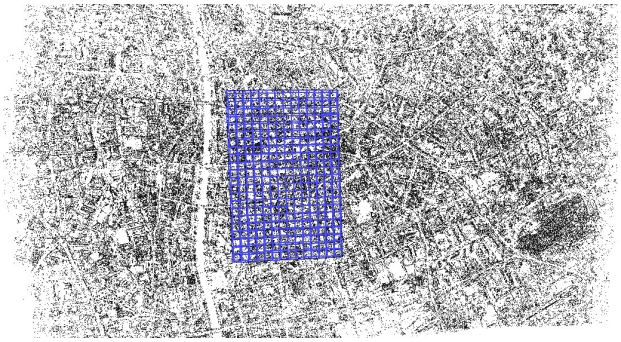


Figure 6: Volumentric multi-view dense matching. A near and far plane parallel to the image plane of the reference camera define the bounding volume.

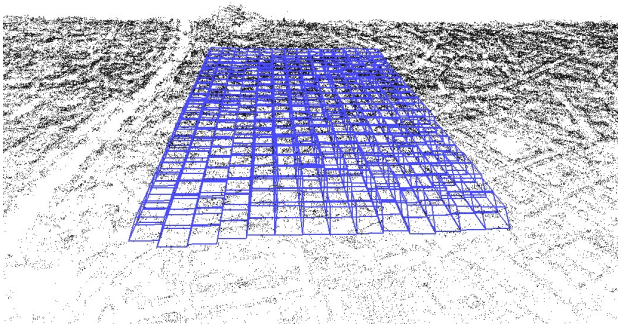
Initial DSM from Public GIS Data We use publicly available elevation data provided by the Shuttle Radar Topography Mission (SRTM) (Farr et al., 2007) to create a coarse approximation of the Earth's surface. This data serves as a polygonal 3D surface model and is used to limit the potential depth range for plane sweep. SRTM provides a digital elevation model (DEM) of the Earth at near-global scale, covering about 80% of the Earth's total landmass. The dataset is available to the public in 2.5D raster format at 1 arc-sec resolution (SRTM-1, approximately 30 meters) over the United States and its territories and at 3 arc-seconds resolution (SRTM-3, approximately 90 meters) for the rest of the world. We combine the DEM with information for buildings from freely available 2D vector map data from the OpenStreetMap[2] (OSM) project. Besides street networks and manifold points of interest (POIs), the OSM project provides outlines of buildings for many cities around the world.

Geometry of the initial DSM is represented as a triangulated irregular network (TIN) (Peucker et al., 1978) of 3D points. Build-

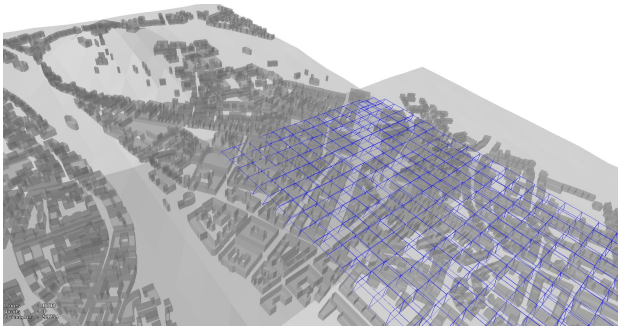
[2] http://www.openstreetmap.org



(a)



(b)



(c)

Figure 7: Tiling and depth range estimation for one specific key view of the *Graz* dataset from sparse points (a),(b) and by DSM approximation from public domain geographic data sources (c).

ings are modeled as polyhedral objects by extruding building footprints to predefined height values for the maximum expected building height (Figure 7). This may also allow dense reconstruction algorithms to take advantage of already known scene geometry for applications such as visibility checks and occlusion handling.

In addition to the scene volume extent, a depth sampling Δd and the number of depth steps in the volume is chosen such that subpixel accurate matching is achieved. The depth step Δd is adaptively computed such that the Nyquist criterion (Shannon, 1949) $f_s > 2$ pixel is satisfied for at least half of all sensor views. This means that for 50% of potential sensor views i, the following condition must be satisfied,

$$\text{median } (||p(P_i, X(d)) - p(P_i, X(d+\Delta d))||) < 0.5 \text{ pixel}, \quad (6)$$

where $X(d)$ is the point passing trough the center of every tile at depth d, P_i is the projection matrix of view i and p the projection operator. This ensures that sampling artifacts are avoided for at least 50% of all sensor views.

Image Correlation We use normalized cross correlation (NCC) as photo consistency measure for plane sweep cost computation. The correlation between two signals (cross-correlation) is a robust approach for dense matching. One advantage of normalized cross correlation (NCC) compared to simpler methods like SAD and SSD is the invariance to linear intensity changes which often occur in aerial images. Given two intensity vectors $I_1 \in \mathcal{R}^n$ and $I_2 \in \mathcal{R}^n$, the normalized cross-correlation is computed by,

$$\rho = \frac{\sum_{k=1}^{n}(I_1(k) - \bar{I}_1)(I_2(k) - \bar{I}_2)}{\sqrt{\sum_{k=1}^{n}(I_1(k) - \bar{I}_1)^2 \sum_{k=1}^{n}(I_2(k) - \bar{I}_2)^2}} \quad (7)$$

where $\bar{I}_1$, $\bar{I}_2$ are the mean intensities and n is the length of the intensity vector. Note that if two image patches match perfectly, the normalized cross-correlation value is 1.

2.3 Cost Aggregation and Implicit Occlusion Handling

In order to handle occlusion that often occur in a multi-view setup, we use truncated correlation measures between the key view and the N sensor views,

$$C(x, d) = \frac{1}{N} \sum_{i=1}^{N} \min(d_W(I(x), I_i(x, d)), t) \quad (8)$$

where x is the pixel position in the key view I, d the current depth and I_i the respective sensor view. The image similarity function d_W is evaluated in a $k \times k$ neighborhood and t is a constant threshold that accounts for occlusions and outliers.

Since NCC delivers correlation values ρ between $[-1 \dots 1]$, the image similarity score (i.e. matching costs) is computed using,

$$d_W(I(x), I_i(x, d)) = \frac{1 - \rho}{2}, \quad (9)$$

where a perfect correlation value implies zero costs.

2.4 Depth Map Extraction

From the 3D cost volume, dense depth maps can be extracted using global optimization methods. Given a graph with node set $\mathcal{V}$, edges $\mathcal{E}$ and a label set $\mathcal{L} \subset \mathcal{Z}$, an optimal labeling $l \in \mathcal{L}^{\mathcal{V}}$ for the energy of the form,

$$\min_l \sum_{(u,v)\in\mathcal{E}} P(l(u) - l(v)) + \sum_{v\in\mathcal{V}} D(l(v)) \quad (10)$$

where $P(l(u) - l(v))$ are pairwise potentials and $D(l(v))$ is the unary term, respectively. Solving this problem corresponds to a minimal cut (Kolmogorov and Zabih, 2002) on a graph in higher dimensions where labels are ordered. In (Ishikawa, 2003) a minimum cut algorithm is presented that exactly solves this class of Markov Random Field (MRF) problem. This problem perfectly fits to dense depth estimation, where $l(v) \in \mathcal{L}$ are depth labels, $v \in \mathcal{V}$ pixels and $\mathcal{E}$ describes the connection of pixels. Such a labeling combines a certain pairwise regularity term $P(\cdot)$ with an arbitrary data term $D(\cdot)$. In (Pock et al., 2008) a continuous formulation to the discrete multi-label problem of Ishikawa is given. The corresponding variational problem to Equation (10) is,

$$\min_u \{ \int_\Omega |\nabla u| + \int_\Omega C(x, u(x)) dx \}, \quad (11)$$

where $u : \Omega \rightarrow \Gamma$ is the unknown function and $\Omega \subseteq \mathcal{R}^2$ is the image domain. $\Gamma = [\gamma_{min}, \gamma_{max}]$ is the range of u. The left term $|\nabla u|$ is the total variation (TV) term that allows for sharp discontinuities in the solution while still being a convex function. This is a desired property for dense matching where edges should be preserved in the solution. The right term of Equation (11) is the data term measuring the matching quality for a given u between the key view and sensor views. The spatial continuous formulation comes along with several advantages over the discrete approach. On the one hand continuous optimization can be implemented using simple and efficient primal-dual optimization techniques which can be easily accelerated on parallel architectures such as graphics processing units (GPUs). On the other hand these methods require considerably less memory which makes the method applicable for quite large practical problems (Pock et al., 2010).

3 RESULTS AND DISCUSSION

We perform dense matching for a sub-block of the aerial dataset *Graz* as shown in Figure 3. For each key view the set of overlapping sensor views is determined. The overlap is computed from sparse correspondences obtained by the aerial triangulation. Only images with an overlap of more than 10% are considered, which means that each key view has about ten overlapping sensor views. Our dense matching algorithm requires a cost volume of size $W \times H \times D$ which depends on the image width W and height H of each image and the number of depth labels D. Since a global cost volume would be too large to fit into GPU memory, the area of interest has to be divided into tiles (e.g. 512×512). Each tile is processed independently, but with a sufficient overlap in order to suppress boundary effects. Figure 7 shows a 512×512 tiling of one specific key view.

For our experiments we set the NCC matching window radius to $r = 1$ pixel and $t = 0.5$ for the outlier and occlusion threshold in the cost accumulation step. A regularization parameter of $\lambda = 20$ is used in the continuous optimization. This parameter balances between data and regularity term and determines the degree of smoothness of the extracted depth maps. Processing of a cost volume of size $512 \times 512 \times 128$ requires about 1.5 minutes on a Nvidia GeForce GTX280. Performance metrics and detailed processing timings for dense matching are summarized in Table 1.

Figure 8 shows depth maps computed by a local winner takes all (WTA) approach and the global multi-label optimization as described in Section 2. While the WTA approach leads to noisy depth maps due to matching ambiguities, the global method produces clean results while still preserving sharp edges at discontinuities. This can be seen from Figure 9 that depicts an oblique view of the textured depth map.

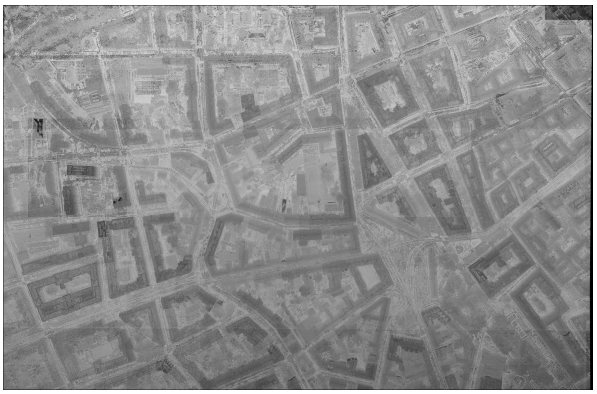
image resolution [pixel]	7500×11500
tile size [pixel]	512×512
number of tiles	384
max number of depths [s]	160
matching time per slice [s]	0.076
global optimization time [s]	74
total time per tile [s]	90

Table 1: Performance metrics and timings for processing one high resolution image on a singe GPU (Nvidia GeForce GTX280).

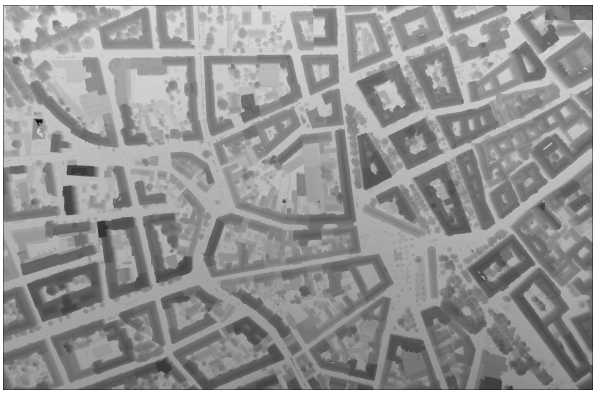
4 CONCLUSION

In this paper we presented an approach for fully automated aerial triangulation and dense matching from large aerial images. The method relies on image data only and does not require any external orientation sensor such as GPS/INS. Hence, the proposed method is very flexible to apply. We present an algorithm for efficient and fully automated aerial dense matching using a multi-view approach based on plane-sweep. A global optimization algorithm based on a continuous energy minimization framework delivers globally optimal solutions with respect to our discontinuity preserving multi-view cost function. We successfully demonstrated that our multi-view matching technique achieves highly accurate dense reconstruction results from large aerial images.

(a)



(b)



(c)

Figure 8: (a) Key image and depth maps produced by multi-view dense matching using winner takes all (b) and continuous multi-label optimization (c).

ACKNOWLEDGEMENTS

This work was supported by the Austrian Research Promotion Agency (FFG) FIT-IT project HOLISTIC (830044).

(a) (b)

Figure 9: Oblique point of view of texturized depth maps from Graz with a GSD of 10cm.

REFERENCES

Baltsavias, E., 1999. A comparison between photogrammetry and laser scanning. ISPRS Journal of Photogrammetry and Remote Sensing 54(2-3), pp. 83–94.

Beder, C. and Steffen, R., 2006. Determining an initial image pair for fixing the scale of a 3d reconstruction from an image sequence. In: Annual Symposium of the German Association for Pattern Recognition (DAGM), pp. 657–666.

Collins, R. T., 1996. A space-sweep approach to true multi-image matching. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 358–363.

Farr, T. G. et al., 2007. The shuttle radar topography mission. Technical report, Rev. Geophys., 45, RG2004.

Fischler, M. A. and Bolles, R. C., 1981. Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. Communication Association and Computing Machine 24(6), pp. 381–395.

Gallup, D., Frahm, J.-M. and Pollefeys, M., 2008. Variable baseline/resolution stereo. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Hirschmüller, H., 2006. Stereo vision in structured environments by consistent semi-global matching. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. II: 2386–2393.

Hutton, J. and Mostafa, M. M., 2005. 10 years of direct goereferencing for airborne photogrammetry. In: Photogrammetric Week.

Ishikawa, H., 2003. Exact optimization for markov random fields with convex priors. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(10), pp. 1333–1336.

Kolmogorov, V. and Zabih, R., 2002. Multi-camera scene reconstruction via graph cuts. In: European Conference on Computer Vision (ECCV), pp. 82–96.

Leberl, F. and Gruber, M., 2003. Flying the new large format digital camera ultracam-d. In: Proceedings of the Photogrammetric Week, Stuttgart University.

Leberl, F., Irschara, A., Pock, T., Meixner, P., Gruber, M., Scholz, S. and Wiechert, A., 2010. Point clouds: Lidar versus 3d vision. Photogrammetric Engineering and Remote Sensing.

Lowe, D., 2004. Distinctive image features from scale-invariant keypoints. Int. Journal of Computer Vision 60(2), pp. 91–110.

Nistér, D., 2004. An efficient solution to the five-point relative pose problem. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 26(6), pp. 756–770.

Nistér, D. and Stewenius, H., 2006. Scalable recognition with a vocabulary tree. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2161–2168.

Peucker, T. K., Fowler, R. J., Little, J. J. and Mark, D. M., 1978. The triangulated irregular network. Amer Soc Photogrammetry Proc Digital Terrain Models Symposium 516, pp. 96–103.

Pock, T., Cremers, D., Bischof, H. and Chambolle, A., 2010. Global solutions of variational models with convex regularization. SIAM J. Imaging Sciences 3(4), pp. 1122–1145.

Pock, T., Schoenemann, T., Graber, G., Bischof, H. and Cremers, D., 2008. A convex formulation of continuous multi-label problems. In: European Conference on Computer Vision (ECCV), Marseille, France.

Rumpler, M., Irschara, A. and Bischof, H., 2011. Multi-view stereo: Redundancy benefits for 3d reconstruction. In: Proceedings of the 35th Workshop of the Austrian Association for Pattern Recognition.

Seitz, S., Curless, B., Diebel, J., Scharstein, D. and Szeliski, R., 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 519–526.

Shannon, C. E., 1949. Communication in the presence of noise. Proceedings of the Institute of Radio Engineers (IRE) / IEEE 37, pp. 10–21.

Sivic, J. and Zisserman, A., 2003. Video google: A text retrieval approach to object matching in videos. In: IEEE International Conference on Computer Vision (ICCV), pp. 1470–1477.

Snavely, N., Seitz, S. M. and Szeliski, R. S., 2008. Modeling the world from internet photo collections. Int. Journal of Computer Vision 80(2), pp. 189–210.

Triggs, B., McLauchlan, P., Hartley, R. and Fitzgibbon, A., 2000. Bundle adjustment – A modern synthesis. In: Vision Algorithms: Theory and Practice, pp. 298–375.

Zebedin, L., 2010. Automatic Reconstruction of Urban Environments from Aerial Images. PhD thesis, Graz University of Technology.