Multi-View Stereo: Redundancy Benefits for 3D Reconstruction

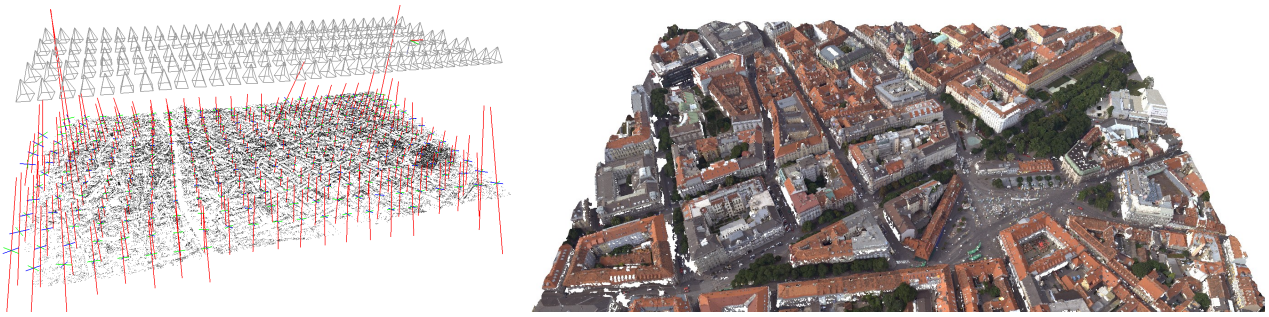
Markus Rumpler, Arnold Irschara and Horst Bischof

Institute for Computer Graphics and Vision
Graz University of Technology, Austria
{*rumpler,irschara,bischof*}*@icg.tugraz.at*

Abstract
This work investigates the influence of using multiple views for 3D reconstruction with respect to depth accuracy and robustness. In particular we show that multiview matching not only contributes to scene completeness, but also improves depth accuracy by improved triangulation angles. We first start by synthetic experiments on a typical aerial photogrammetric camera network and investigate how baseline (i.e. triangulation angle) and redundancy affect the depth error. Our evaluation also includes a comparison between combined pairwise triangulated and fused stereo pairs in contrast to true multiview triangulation. By analyzing the 3D uncertainty ellipsoid of triangulated points we demonstrate the clear advantage of a multiview approach over fused two view stereo algorithms. We propose an efficient dense matching algorithm that utilizes pairwise optical flow followed by a robust correspondence chaining approach. We provide evaluation results of the proposed method on ground truth data and compare its performance in contrast to a multiview plane sweep method.

1. Introduction

Image-based 3D reconstruction is an active field of research in Photogrammetry and Computer Vision. The need for detailed 3D models for mapping and navigation, inspection, cultural heritage conservation or photorealistic image-based rendering for the entertainment industry lead to the development of several techniques to recover the shape of objects. To achieve precise and high detailed reconstructions *Lidar* is often employed providing 2.5D range images and the respective 3D point cloud in a metric scale. On the other hand, laser-based methods are complex to handle for large scale outdoor scenes, especially for aerial data acquisition. In contrast to that, passive image-based methods that utilize multiple overlapping views are easily deployable and are low cost compared to Lidar [7], but require some post-processing effort to derive depth information. In this work we investigate how redundancy and baseline influence the depth accuracy of multiple view matching methods. In particular we perform synthetic experiments on a typical aerial camera network that corresponds to a 2D flight pattern with 80% forward-overlap and 60% side-lap (see Figure 1). By covariance analysis of triangulated scene points [1, 2], the theoretical bound of depth accuracy is determined according to the triangulation angle and the number of measurements (i.e. the redundancy). One of our main findings is that true multiview matching/triangulation outperforms two-view fused stereo results by at least one order of magnitude in terms of depth accuracy. Furthermore, we present a fast, accurate and robust multiview matching and reconstruction technique suitable for high resolution images of large scale scenes that is able to compete with Lidar through leveraging the redundancy of many views. Our proposed solution to multiview reconstruction is based on pair-wise stereo, employing efficient and robust TV-L^1 [14] optical flow that is restricted to the epipolar geometry. Unlike standard aerial



(a) (b)

Figure 1. Multi-view 3D reconstruction from aerial images. (a) The view network, a sparse reconstruction and uncertainties (magnified by 1000 for better visibility) for selected 3D points on a regularly sampled grid on the ground plane. (b) Reconstructed dense point cloud from our multiview method of an urban scene.

matching approaches that rely on 2.5D data fusion [16] of pairwise stereo depth maps, we propose a correspondence chaining (i.e. measurement linking) and triangulation approach that takes full advantage of the achievable baseline (i.e triangulation angles). In contrast to *voxel-based* approaches [13], *polygonal meshes* [6] and *local patches* [3], we focus on algorithms representing geometry as a set of *depth maps* [15]. It eliminates the need for resampling the geometry in the three-dimensional domain and can be easily parallelized. We evaluate our approach on the multiview benchmark dataset of Strecha et al. [12] that provides accurate ground truth and on large scale aerial images.

2. Uncertainty of Scene Points

As shown in [4] the depth uncertainty of a rectified stereo pair can be directly determined from the disparity error,

$$\epsilon_z = \frac{bf}{d} - \frac{bf}{d+\epsilon_d} \approx \frac{z^2}{bf} \cdot \epsilon_d \tag{1}$$

where z is the point depth, f the focal length and b the image baseline. Hence, the depth precision is mainly a function of the ray intersection angle. In contrast, for multiview image matching and triangulation, the redundancy not only implies more measurements but additionally constrains the 3D point location through multiple ray intersections. These entities are not independent but are coupled, since they rely on the network geometric configuration that determines image overlap (i.e. redundancy) and baseline, simultaneously. 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 $C_{\mathbf{X}}$) as shown in [1]. An empirical estimate of the covariance ellipsoid corresponding to multiview triangulation can be computed by statistical simulation. For the moment we assume that camera orientations and 3D structure are fixed and known. The cameras are distributed along a 2D grid (corresponding to flight paths) in order to achieve a 80% forward overlap and 60% side-lap (see Figure 1). According to a large format digital aerial camera (e.g. UltraCamD from Microsoft) the image resolution is set to 7500×11500 pixel with a field of view $\alpha = 54°$. Furthermore, 3D points are evenly distributed on a 2D plane that corresponds to the bold earth surface, observed from a flying height of 900m. Therefore, an average Ground Sampling Distance (GSD) of 8cm/pixel is achieved.

Given the cameras $P_{i=1:N} \in \mathcal{P}$ (i.e. calibration and poses) and 3D points $\mathbf{X}_{j=1:M} \in \mathcal{X}$, respective ground truth projections are produced $x_{ij} = P\mathbf{X}$. Therefore, for every 3D point a set of point-tracks

(i.e. 2D measurements) is generated $m = (< x_1, y_1 >, < x_2, y_2 > \ldots, < x_k, y_k >)$. Next, 2D projections are perturbed by zero mean Gaussian isotropic noise $\hat{x} = x + \mathcal{N}(0, \Sigma)$,

$$\Sigma = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix} \tag{2}$$

with standard deviation $\sigma_x = \sigma_y = 1$ pixel (i.e. $\sim$ 8cm GSD). Given the set of perturbed point tracks $\hat{m} = (< \hat{x}_1, \hat{y}_1 >, < \hat{x}_2, \hat{y}_2 > \ldots, < \hat{x}_k, \hat{y}_k >)$ and ground truth projection matrices $P_{i=1:N}$, the 3D position of the respective point in space is determined. This process requires the intersection of at least two known rays in space. Hence, we use a linear triangulation method [5] to determine the 3D position of point tracks. This method generalizes easily to the intersection of multiple rays providing a least-squares solution. Optionally, a non-linear optimizer based on the Levenberg-Marquardt algorithm is used to refine the 3D point by minimizing the reprojection error. Through Monte Carlo Simulation on the perturbed measurement vectors $\hat{m}$ we obtain a distribution of 3D points $\mathbf{X}_i$ around a mean position $\hat{\mathbf{X}}$. From the Law of Large Numbers it follows that for a large number N of simulations, one can approximate the mean 3D position by,

$$E_N[\mathbf{X}_i] = \frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i \tag{3}$$

and its respective covariance matrix by,

$$C_{\mathbf{X}} = E_N[(\mathbf{X}_i - E_N[\mathbf{X}_i])(\mathbf{X}_i - E_N[\mathbf{X}_i])^\top] \tag{4}$$

Using the singular value decomposition the covariance matrix can then be diagonalized,

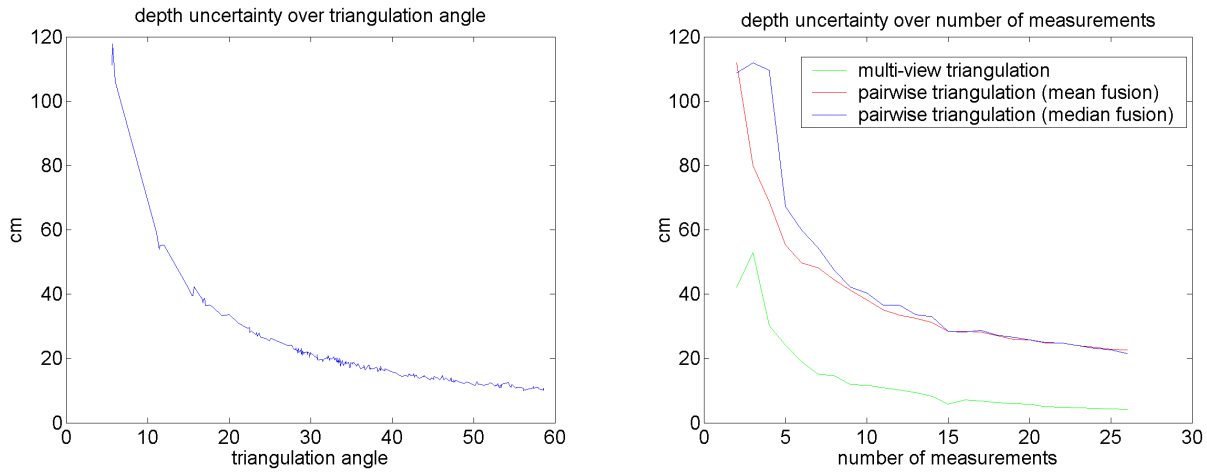
$$C_{\mathbf{X}} = U \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix} V^\top \tag{5}$$

where U represents the main diagonals of the covariance ellipsoid and σ_i are the respective standard deviations. The decomposition of the covariance matrix (equation 5) into its main diagonals directly relates to the uncertainty in $x - y$ and z direction. Under the assumption of fronto-parallel image acquisition the largest singular value σ_1 corresponds to the uncertainty in depth and σ_2 and σ_3 to the uncertainty in $x - y$ direction, respectively.

3. Synthetic Simulation

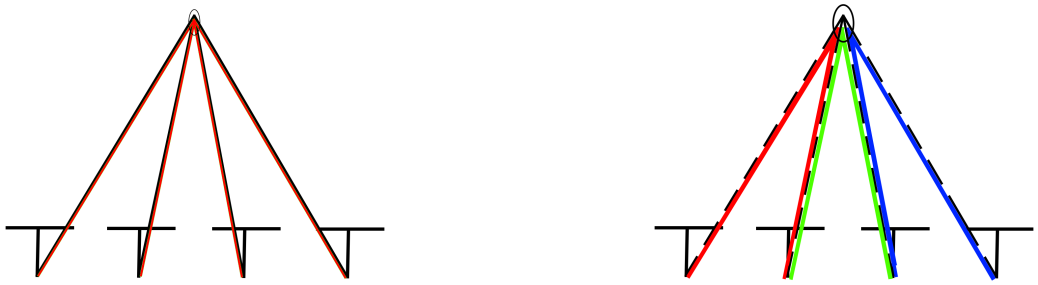
For the aerial network described in Section 2. we perform synthetic experiments where we investigate the influence of baseline (i.e. triangulation angle) and the number of views used for matching/triangulation (i.e. redundancy).

Triangulation angle. In our first experiment we consider the depth error (i.e. the uncertainty along the z direction) as a function of triangulation angle. To this end we randomly choose homologous points from camera pairs $< P_i, P_j >$, $i \neq j$ and compute the depth error by decomposing the covariance matrix according equation 5. Figure 2a depicts the depth error σ_z (i.e. uncertainty along the z axis) with respect to the achieved triangulation angle,



(a) (b)

Figure 2. (a) Depth uncertainty decreases with larger triangulation angles. (b) Comparison of depth uncertainty for fused 3D points from pairwise triangulation in contrast to true multiview triangulation with respect to the number of measurements.



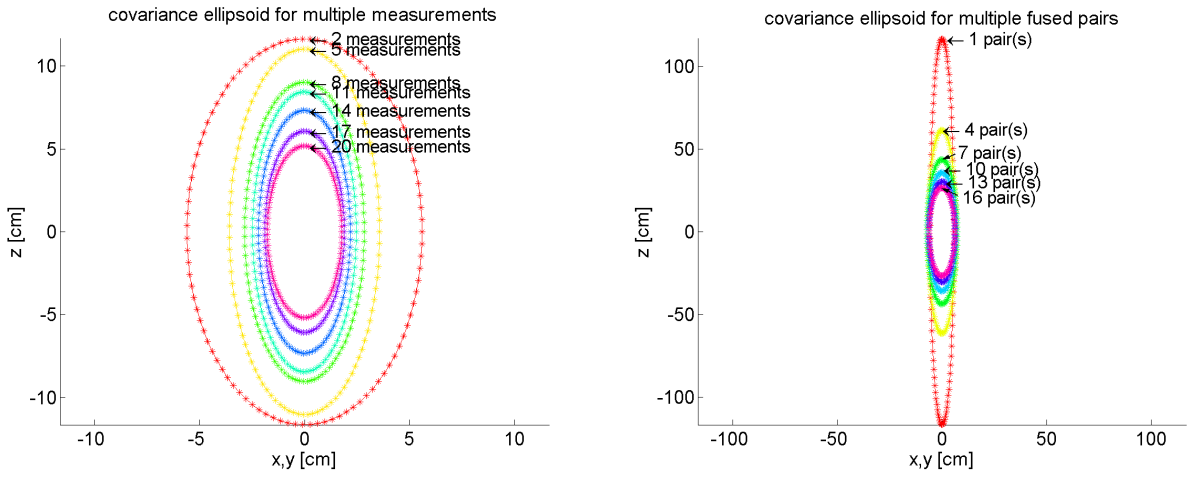
(a) (b)

Figure 3. (a) True multiview triangulation. (b) Pairwise triangulation and pairwise 2.5D stereo fusion.

$$\alpha = \arccos \frac{\overline{\mathbf{C}_i\mathbf{X}} \cdot \overline{\mathbf{C}_j\mathbf{X}}}{||\overline{\mathbf{C}_i\mathbf{X}}||||\overline{\mathbf{C}_j\mathbf{X}}||} \tag{6}$$

where $\mathbf{X}$ is the triangulated 3D point and $\mathbf{C}_i$, $\mathbf{C}_j$ are the camera centers of the first and second camera, respectively. A narrow triangulation angle of 10° translates into an absolute depth uncertainty $\sigma_z \approx 70$cm, whereas angles of $> 50^\circ$ produce a depth error of less than 15cm.

Multi-view triangulation. In our second experiment we investigate the influence of redundancy with respect to the depth error. The geometric configuration of a camera network implies that redundancy and triangulation angle cannot be decoupled since adding more views will result in a set of different triangulation angles. To compensate for this effect, we first determine the image pair that minimizes the depth error (i.e. the pair with maximum triangulation angle). Next, additional measurements from neighboring views are added randomly one by one and the depth error is computed. Figure 4a depicts the covariance along the (x, z)-plane for one 3D point with respect to the number of measurements used for triangulation. Note, while the shape of the uncertainty ellipsoid varies with increasing number of measurements, the overall accuracy along each axis decreases. This is in contrast to the methods proposed by Beder et al. [1] that considers the roundness of the uncertainty ellipsoid $R = \sqrt{\frac{\sigma_3^2}{\sigma_1^2}}$ as uncertainty measure.



(a) (b)

Figure 4. Covariance ellipsoids for one exemplary 3D point. (a) Starting from the image pair that minimizes the depth error (i.e. the pair with maximum angle), additional measurements are added one by one for multiview triangulation. (b) Covariance ellipsoid for a varying number of fused stereo pairs of adjacent views (cf. Fig. 3b).

Stereo triangulation fusion. We compare the multiview triangulation result with those of fused pairwise triangulated stereo pairs (see Figure 3). Stereo pairs $(< P_1, P_2 >, < P_2, P_3 > \ldots, < P_{k-1}, P_k >)$ are selected from consecutive views $< P_i, P_{i+1} >$ for $i = 1 \ldots k-1$ along the flight path. Each pair is used for triangulation of $k-1$ 3D points $\mathbf{X}_{<\mathbf{i},\mathbf{i+1}>}$ that belong to one and the same point-track m. The mean and median value of this set is determined representing the fused depth estimate. The covariance ellipsoids corresponding to the uncertainty of one exemplary 3D point fused from a varying number of stereo pairs is depicted in Figure 4b. Note, by using more stereo pairs, the uncertainty in depth decreases but overall the fused stereo result cannot compete with multiview triangulation/matching (Figure 4a). For instance, while the uncertainty of 16 fused stereo pairs gives a depth error $\sigma_z \approx 25$cm, the multiview-triangulation leads to a $\sigma_z \approx 6$cm. A comparison of multiview vs. fused stereo triangulation uncertainty with respect to the number of used measurements/views is depicted in Figure 2b.

4. Multi-View Depth Maps

Based on the synthetic experiments of the previous section we conclude that true multiview image matching and triangulation offers clear advantages in terms of depth accuracy over fusion of multiple stereo pairs. We propose a dense matching approach that considers only pairs of images, thus being fast and efficient, but leverages multiple views by linking measurements between view pairs (i.e. correspondence chaining), thus increasing the baseline/triangulation angle.

4.1. Stereo Matching

Our stereo dense matching approach is based on TV-L^1 optical flow [14] computation. *Optical flow* seeks to determine a displacement field u between two images I_0 and I_1 estimating the motion of pixels, i.e. the mapping of image points from the first image to their new location in the second one. Hence, optical flow is equivalent to the search for correspondences in stereo vision. Disparities between pixels are estimated within a global optimization framework that seeks a solution by minimizing an appropriate energy function. The approach is based on *total variation* (TV) [10] regu-

larization and uses a robust L^1 data fidelity term. This allows for smoothness while preserving depth discontinuities to obtain high quality depth maps. The known camera parameters imply that the correspondence search can be restricted to one dimension along the epipolar line [11]. The computational effort is then similar to a standard stereo case with a rectified image pair. The epipolar line in the sensor view for a key view pixel x is given with $x \mapsto l' : l' = Fx$. The direction of the epipolar line given by the unit vector l'_n together with a point on the line (i.e. an initial reference point obtained from SfM) with a given disparity u_0 yields to the location of the point correspondence x': $x' = x_{ref} + u_0 l'_n$. Hence, we linearize image I_1 near x'. With I_1^e denoting the derivative with respect to the epipolar direction, the energy functional writes,

$$E = \int_\Omega \{\lambda |u I_1^e + I_1(x') - u_0 I_1^e - I_0| + |\nabla u|\}\, dx. \tag{7}$$

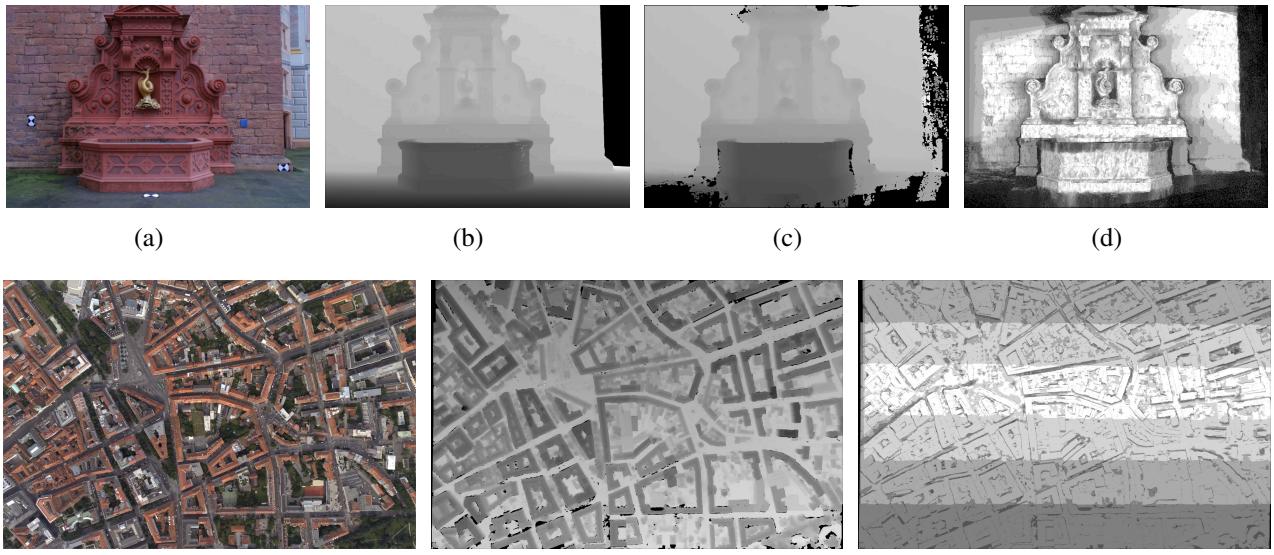
This energy can be efficiently minimized by a primal dual algorithm [14].

4.2. Correspondence Chaining

Dense correspondence computation is performed between pairs of images as described in the previous section, but is not restricted to that specific method. A pair always consists of the key view and one of its neighboring sensor views. We seek for a set of correspondences (i.e. measurements) for each pixel of the key view, one from every neighboring view in which the pixel is visible. Reliable correspondence estimates from optical flow can be expected only for adjacent neighbors since wide baseline settings normally result in larger distortions and occlusions, thus degrading matching confidence and accuracy [8]. On the other hand, small baselines introduce inaccuracies due to narrow triangulation angles (see Section 3). To overcome these problems, we follow a multi-baseline approach [4] and propose *correspondence chaining* to enhance baseline (i.e. triangulation angle) and redundancy at the same time. Starting from the adjacent neighbors $P_{k\pm1}$ of each key view P_k, we chain flow vectors from one view to the next. If a disparity estimate $u_{l,c}$ between the linking view l and the next sensor view c is available, we update the coordinates of the measurement according to $x'_c = x_k + u_{k,l}(x_k) + u_{l,c}(x_k + u_{k,l}(x_k)) = x'_l + u_{l,c}(x'_l)$. This approach can fail, if a wrong correspondence (i.e. an outlier) is used for linking. The error propagates over all links and corrupts the correct depth estimation at that pixel. Therefore, we employ an outlier rejection strategy based on the *RANdom SAmple Consensus* (RANSAC) [5] algorithm to provide robust depth estimates in the reconstruction. From the set of at least three inliers, a least squares solution is computed.

4.3. Experimental Results

We evaluate our correspondence chaining approach on large scale aerial images (see Figure 1) and on the multiview dataset from Strecha et al. [12] that provides ground truth data for a quantitative evaluation (Figure 5). We compare our proposed approach to a multiview plane sweep method that implicitly combines multiple measurements through a three dimensional voxel space of truncated matching costs. Three different photoconsistency measures are used: sum of absolute differences (SAD), zero mean normalized sum of absolute differences (ZNSAD) and zero mean normalized cross correlation (ZNCC). Furthermore, a global optimization method [9] is used for robust and smooth disparity assignment. We calculate the root mean square (RMS) error measured in depth units between the ground truth d_r and the depth map d_c, $RMS = \sqrt{\frac{1}{N}\sum_{(x,y)} |d_r(x,y) - d_c(x,y)|^2}$. In addition, we compute the completeness of the scene, a measure that determines the percentage of estimated depth values with respect to the total number of pixels available in the reference maps. Table 1 summarizes



(a) (b) (c) (d)

(e) (f) (g)

Figure 5. (a) Image from the fountain-P11 dataset [12]. (b) Reference depth map from ground truth. (c) Result from our TV-L^1 stereo based multiview method. (d) Confidence map depicting the number of used inliers for triangulation from correspondence chaining. Low image overlap and occlusions are clearly visible (dark areas). (e) Key image from an aerial dataset, (f) depth map and (g) inlier confidence map.

		flow	plane sweep		
			SAD	ZNSAD	ZNCC
fountain-P11	RMS error	0.257	0.71454	0.540	0.421878
	completeness [%]	93.055	94.7247	94.658	94.6586

Table 1. Error statistics for image 5 of the fountain-P11 dataset [12]. 11 views (10 pairs) are used for correspondence chaining. The flow reconstruction method was initialized with a depth estimate from the sparse points. Parameters for TV-L^1 matching: $\lambda = 0.15$, warps=5, iterations=100. Parameters for plane sweep and global optimization: $\lambda = 100, t = 0.17$ (SAD, ZNSAD) and $\lambda = 20, t = 0.5$ (ZNCC).

our evaluation. From our experiments we conclude that our proposed approach compares well to the multiview plane sweep method in terms of accuracy and completeness.

5. Conclusion

In this paper we analyzed and evaluated multiple view matching methods with respect to baseline and redundancy. From our synthetic experiments we conclude that true multiview matching/triangulation outperforms two-view stereo approaches by about one order of magnitude. Furthermore, we presented a fast and accurate multiview matching method based on TV-L^1 stereo and robust flow chaining that leverages redundancy of multiple views and outperforms current multiview plane sweep approaches.

Acknowledgement

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

References

[1] Christian Beder and Richard Steffen. Determining an initial image pair for fixing the scale of a 3d reconstruction from an image sequence. In *Proc. DAGM*, pages 657–666, 2006.

[2] Wolfgang Förstner. Uncertainty and Projective Geometry. In *Handbook of Computational Geometry for Pattern Recognition, Computer Vision, Neurocomputing and Robotics*. Springer, 2005.

[3] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1362–1376, 2010.

[4] David Gallup, Jan-Michael Frahm, and Marc Pollefeys. Variable baseline/resolution stereo. In *CVPR*, 2008.

[5] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[6] Vu Hoang Hiep, Renaud Keriven, Patrick Labatut, and Jean-Philippe Pons. Towards high-resolution large-scale multi-view stereo. In *CVPR*, pages 1430–1437, 2009.

[7] Franz Leberl, Arnold Irschara, Thomas Pock, Philipp Meixner, et al. Point clouds: Lidar versus 3d vision. *Photogrammetric Engineering and Remote Sensing*, 2010.

[8] Masatoshi Okutomi and Takeo Kanade. A multiple-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:353–363, 1993.

[9] Thomas Pock, Thomas Schoenemann, Gottfried Graber, Horst Bischof, and Daniel Cremers. A convex formulation of continuous multi-label problems. In *ECCV*, pages 792–805, 2008.

[10] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60:259–268, 1992.

[11] Natalia Slesareva, Andres Bruhn, and Joachim Weickert. Optic flow goes stereo: A variational method for estimating discontinuity-preserving dense disparity maps. In *DAGM*, 2005.

[12] Christoph Strecha, Wolfgang von Hansen, Luc Van Gool, Pascal Fua, and Ulrich Thoennessen. On Benchmarking Camera Calibration and Multi-View Stereo for High Resolution Imagery. In *CVPR*, 2008.

[13] George Vogiatzis, Carlos Hernández Esteban, Philip H. S. Torr, and Roberto Cipolla. Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *In PAMI*, 2007.

[14] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l1 optical flow. In *DAGM*, pages 214–223, 2007.

[15] Christopher Zach, Mario Sormann, and Konrad Karner. High-performance multi-view reconstruction. In *3DPVT*, 2006.

[16] Lukas Zebedin, Joachim Bauer, Konrad Karner, and Horst Bischof. Fusion of feature- and area-based information for urban buildings modeling from aerial imagery. In *ECCV*. 2008.