Sparse 3D Reconstruction by Edgel Sweeping

Joachim Bauer, Andreas Klaus, Mario Sormann, Konrad Karner

VRVis Research Center

Inffeldgasse 16, Graz, Austria

e-mail: {bauer, klaus, sormann, karner}@vrvis.at

Abstract

This paper introduces a method for the fast generation of sparse 3D point clouds from multiple oriented images. We use a plane sweeping scheme to compute the 3D location of edge features. A two step approach is used to find a set of tentative hypotheses, which are then refined in an optimization pass. A robust image-based similarity measure is used to verify the 3D hypotheses and identify false positives. We performed experiments on a synthetic data set and on several real datasets.

1 Introduction

The computation of 3D structure from multiple images is one of the most important tasks in computer vision. In literature many different approaches have been described. Originally the various image matching methods were formulated for a stereo image pair. Recent methods [5], incorporate multiple images to achieve a more robust matching result. Correlation-based dense stereo matching methods are usually restricted to small baseline setups or in the case of video-based stereo to image sequences with large overlap. A new method for dense multi-view matching was proposed by [10]. This PDE-based approach is useful for the reconstruction of widely separated views of an object. Dense image matching methods provide a disparity map for the scene, i.e. for every pixel a disparity vector $d(x, y)$ is given. The disparity is often constrained by a smoothness criterion and the ordering constraint. These approaches produce a dense 3D reconstruction, i.e. a 3D coordinate for every image pixel. Another branch are the various voxel coloring methods [6, 9]. Those methods produce a volumetric model of the scene and also work for scenes where the ordering constraint is violated. All the above mentioned reconstruction approaches have the drawback of processing times in the order of minutes. For model-based reconstruction a sparse 3D point cloud is sufficient for the extraction of 3D lines and fitting planes or other 3D primitives. The sparse 3D data also yield robust seed points for a subsequent dense reconstruction.

We present a method inspired by Jung et al. [4] which computes sparse 3D point clouds from multiple oriented images. While the original method is used for reconstructing buildings from aerial images we apply the method for close range scenes. Due to the more challenging

nature of close range scenes, for example the significantly higher depth range, we also incorporate an image-based similarity measure to verify the resulting 3D hypotheses. We apply a plane sweeping scheme to generate 3D hypotheses and verify the hypotheses using purely geometric criteria. In section 2 we outline our approach, in section 3 we present experimental results for synthetic and real image data as well. Section 4 concludes the paper with an outlook on future work.

2 Our Method

A plane sweeping method is used to traverse the volume for which the reconstruction is searched. Figure 1 shows the top view of the setup. Multiple cameras (visualized as triangles) view an object (light gray). One camera is chosen as key camera from which rays are intersected with the 3D sweeping plane. The resulting 3D point is projected into all other slave-cameras. A proximity or similarity criterion is now used to detect tentative 3D hypotheses. In the case of voxel coloring this is the color consistency. In our case it is a combination of proximity and gradient direction. A good hypothesis is characterized by a low re-projection error, i.e. the 2D location of the projected 3D point is close to a feature point in all or many of the slave-cameras. We use edgels as feature points, since edgels can be extracted with sub-pixel accuracy with Canny's method [2].

Every edgel in the key image now defines ray. This ray is intersected with the sweeping plane. In order to accept a hypothesis the re-projection of the 3D intersection point must lie close to an edgel in most of the slave-images. We chose distance thresholds in the range of $0.4 \ldots 1$ pixel. All hypotheses resulting from this sweeping process are subsequently refined and a significant number of false positives is eliminated by enforcing a simple gradient direction constraint. The remaining hypotheses are verified using a fast image-based similarity measure inspired by Lowe's SIFT-features [7]. The similarity measure is based on a histogram comparison and therefore suitable for fast outlier detection.

2.1 Hypotheses Selection

In the sweeping stage we select hypotheses only with the proximity criterion. In the original paper of Jung et al. [4] the proximity criterion is evaluated using a quad-tree for the nearest neighbor search. The quad-tree approach however has a computational complexity of $O(log\ n)$, where n is the number of points in the set, per nearest neighbor query. A more efficient method is the use of distance transforms [1] of the edge points. The distance transform allows us distance queries for 2D point sets in $O(1)$. In our case the distance transform is computed for the edgel locations. Using Chamfer filtering [1] a distance image is derived from the input point set. The distance transforms are calculated for the edgel points of every slave image. The determination of the re-projection error of a 3D hypothesis using the distance transform can then be achieved in $O(1)$. Since the 2D coordinates of the re-projection are non-integers we perform the image access operation in the distance transform images with bi-cubic interpolation. An experiment where the distance transform approach was compared

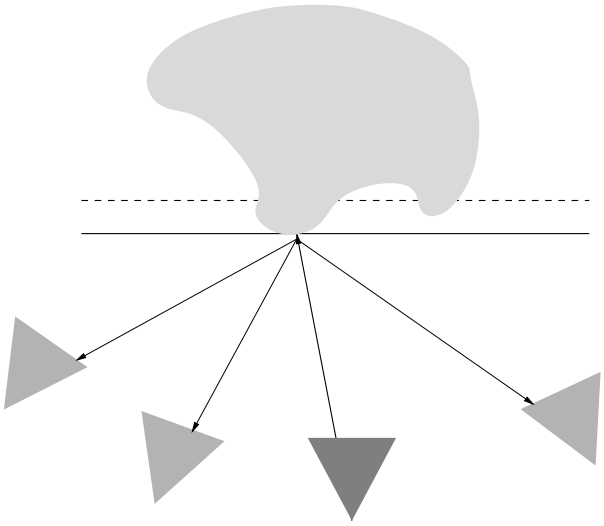


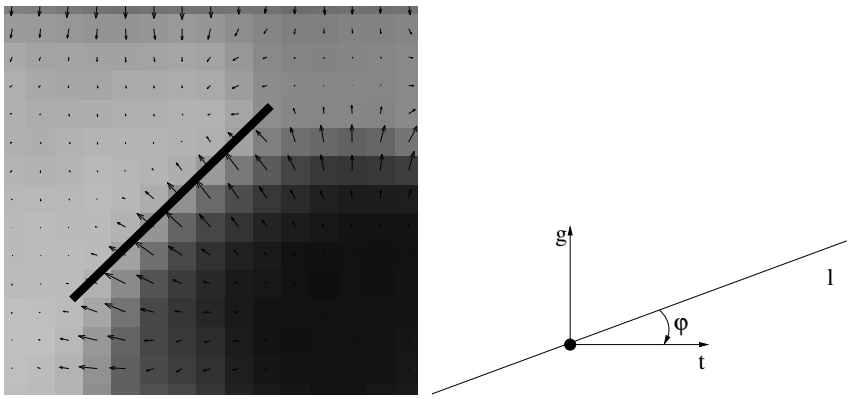
Figure 1: Top view of the setup: Four cameras view an object (light gray). A key camera (dark gray) is chosen and from this camera rays (shown as fat vector) are intersected with the sweeping plane (fat line). The resulting 3D point is projected into all other cameras. The sweep plane moves from front to back, the subsequent instance of the sweep plane is illustrated as dashed line.

with a KD-tree-based nearest neighbor search showed that an average distance error of less than 0.3 pixel can be achieved. This shows that it is justified to use the much faster distance transforms to generate an initial set of 3D hypotheses.

2.2 Hypotheses Refinement

Given the 3D hypotheses that survived the sweeping process, a subsequent refinement optimizes the 3D position and rejects hypotheses that violate the epipolar constraint and a simple gradient direction constraint. For the refinement KD-trees are computed for the edgel sets, since the KD-trees allow accurate nearest neighbor queries and access to additionally stored features such as the gradient vectors. With the epipolar constraint we reject candidates that result from edgels with a gradient direction nearly orthogonal to the direction of the epipolar line. Figure 2 shows a section of an image overlaid with arrows for the gradient directions. Edgels where the gradient direction is nearly orthogonal to the epipolar line (shown in red) are not used for the computation of 3D hypotheses. This strategy avoids ambiguous hypothesis for edgels where the tangent direction (which is orthogonal to the gradient) is nearly parallel to the epipolar line vector.

Another criterion is formed by using the gradient direction g (see Figure 2): the enclosed angle between the gradient direction of the edgel in the key-image and the candidate edgels int the slave images must not exceed a certain threshold. This ensures that all edgels have the same light-to-dark or dark-to-light transition, given that no severe rotation between the images is present.



(a) (b)

Figure 2: Illustration of the epipolar constraint. (a) Section of an image where the edge directions are shown as arrows and the length of the arrows corresponds to the magnitude. (b) Geometry of edgels and epipolar line: the edgel is represented by its gradient direction g and the tangent vector t. Edgels where the tangent vector t is nearly parallel to the epipolar line l are not used for the reconstruction.

Hypotheses that survive the epipolar test are then refined by a fine search in the space around the sweep plane location. Figure 3(a) shows the refinement search space between three instances of the sweep plane (p_1, p_2, p_3). The 3D hypothesis (shown as black dot) is moved along the sweeping direction and projected into all images to evaluate the re-projection error. A possible re-projection error function is shown in the bottom part of Figure 3(a). Only those hypotheses that have a minimum within the search bounds are kept, if the minimum is on the boundaries of the search range, the hypothesis is discarded. Figure 3(b) shows the evaluation of the re-projection error: The goal is to minimize the perpendicular distance d_t from the tangent of the edgel to the projected 3D hypothesis. We evaluate d_t only for the edgel with the smallest Euclidean distance d_e to the projected 3D hypothesis. During the refined search a large amount of the 3D hypotheses is discarded due to non-fulfillment of one of the above criteria.

However, since the search range for terrestrial modeling is significantly higher as for the aerial modeling in the original approach, a number of false positives is still present after the refinement step. In order to detect and remove these false positive matches, a fast image-based similarity measure is applied.

2.3 Image-based Outlier Removal

The similarity measure that is used to eliminate false positive is inspired by Lowe's SIFT-features [7]. This is a scale invariant descriptor which is originally computed for interest points. We want to compute the descriptor for the two rectified image regions that are divided along the edgel tangent direction. Since the SIFT-descriptor is not rotation invariant we need

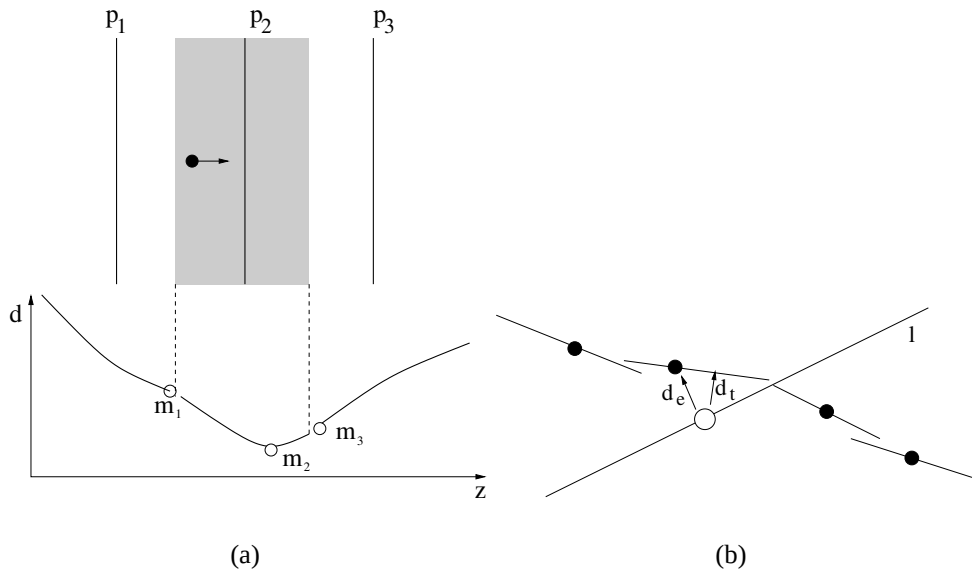


Figure 3: Refinement of 3D hypotheses. (a) shows the search space for the refinement (in light gray) between three instances of the sweep plane (p_1, p_2, p_3). The score function is depicted in the bottom part of (a): only hypotheses that have a minimum within the search boundaries are accepted. This holds only for the minimum m_2 computed for sweep the plane instance p_2, m_1 and m_3 lie on the search range boundaries of the sweep plane instances p_1 and p_3. (b) shows the evaluation of the re-projection error in the 2D edgel sets: the projection of the 3D hypothesis (shown as ring) moves along the epipolar line l. The re-projection error is the perpendicular distance d_t to the tangent of the closest edgel (minimal d_e from the projected hypothesis).

to compute it on rectified frames. Figure 4 shows the rectification approach: On the left side is the original image with an edgel (white dot), its associated tangent direction (white line) and the two attached image regions (shown in different hatching). We use the tangent direction to compute a rectified frame, i.e. we apply an affine transform to align the tangent parallel to the x-axis. The separation of the image area around a candidate edgel is necessary, since edgels often lie on depth discontinuities. Due to these depth discontinuities one of the two image regions that are divided by the edgel tangent may be occluded. Therefore only one side can be used for reliable similarity comparison.

We first calculate the edge orientation φ and magnitude m at each pixel inside the rectified frame I:

$$m(x,y) = \sqrt{(I_{x-1,y} + I_{x+1,y})^2 + (I_{x-1,y} + I_{x+1,y})^2} \tag{1}$$

$$\varphi(x,y) = atan((I_{x-1,y} + I_{x+1,y})/(I_{x-1,y} + I_{x+1,y})). \tag{2}$$

An orientation histogram is used as a region descriptor, the magnitude and the distance of the

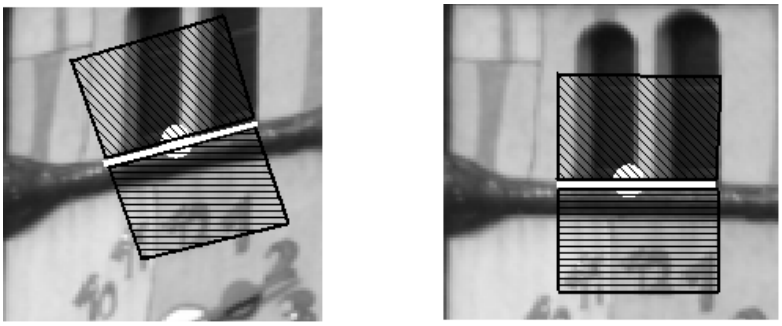


Figure 4: Computing a rectified frame: the tangent direction is used to compute a frame that is aligned parallel to the x-axis.

pixels from the origin are used as a weight. More formally the histogram is calculated as

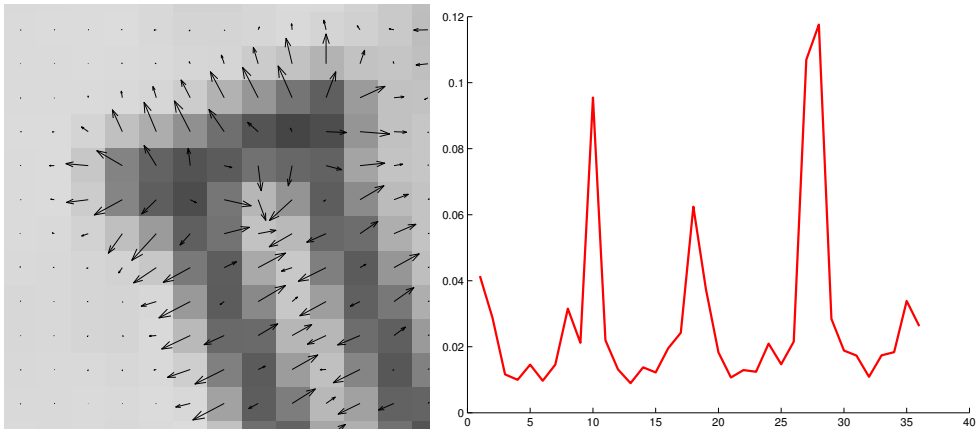
$$H(\theta) = \sum_{\varphi_\epsilon \mathcal{N}} \delta(\theta, \varphi) * w_\varphi, \tag{3}$$

where $H(\theta)$ is the value for bin θ ($\theta \in [0°, 1° \ldots 360°]$) and φ denotes angle values in a neighborhood $\mathcal{N}$ inside the Frame, w_φ is the weight of φ and $\delta(\theta, \varphi)$ is the Kronecker delta function. The angles φ are quantized in accordance with the histogram bins θ. The weight w_φ is computed from the magnitude of φ and a function decreasing with increasing radius r from the origin (x_0, y_0). We use a Gaussian function thus $w_\varphi(x, y) = m(x, y) * g(r)$, with $r = \sqrt{(x - x_0)^2 + (y - y_0)^2}$ and $g(r) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-r^2}{2\sigma^2}}$.

Figure 5(a) visualizes the edge orientations for an image section and Figure 5(b) shows a histogram computed from edge orientations. Since the descriptors are histograms we use a probabilistic distance measure to describe the similarity. Distance measures for histogram comparison are the L_1 and L_2 norm, the Bhattacharyya distance, and the Matusita distance. The earth movers distance is a more complex method for histogram comparison and is computed by solving the so called transportation problem, proposed for image indexing by Rubner et al. [8]. Huet and Hancock [3] give a comparison of the performance of this measures for histogram comparison. Following the conclusions of Rubner we chose the Bhattacharyya distance which is defined as:

$$D_{Bhatt}(H_A, H_B) = -ln \sum_i \sqrt{H_A(i) \cdot H_A(i)} \tag{4}$$

We compute the Bhattacharyya distance separately for the two image regions between the key-image and the slave images. The 3D hypothesis with a distance below a given threshold is finally accepted, if more hypotheses are lying on the same 3D ray, only the one with the smallest distance is accepted.



(a) (b)

Figure 5: Visualization of orientations in the rectified frame: (a) image region with vectors visualizing the edge orientation (vector length corresponds to the magnitude). (b) histogram of edge orientations.

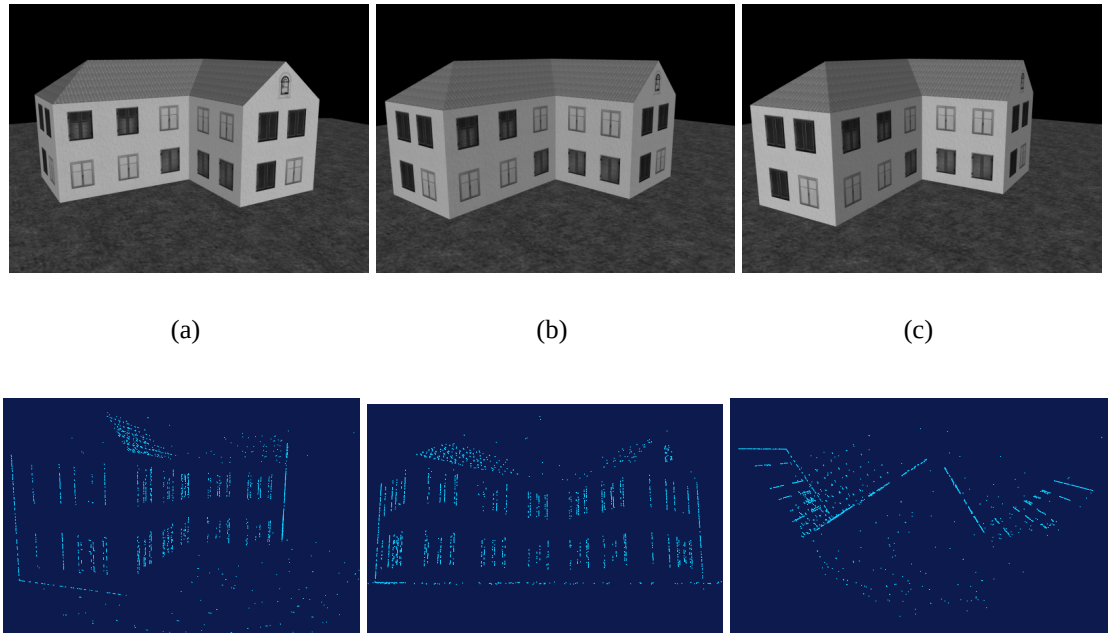
Scene	initial hypotheses	hypotheses remaining after refinement	final 3D points	execution time
synthetic sequence	301532	17971	7142	47 sec.
courtyard sequence	234101	23312	18714	81 sec.
statue sequence	587687	49817	12423	143 sec.

Table 1: Evaluation of the reconstruction performance. Results are given for the 3 image sequences in Figures 7 and 8.

3 Experiments

Several experiments were carried out on a synthetic and real data. Figures 6, 7 and 8 show one image of each scene used. Every scene consists of five images and the threshold for the re-projection error is set to 0.4 pixel. The size of the support region for the image-based verification of the refined hypotheses was set to 15×15 pixel.

The interior camera parameters are determined by an offline camera calibration and the exterior camera orientation parameters of the images are determined by automatic multi-image matching followed by an estimation of the relative orientation of the sequence and a final bundle adjustment. The depth range for the sweeping is ten times the average baseline between the cameras. Only a small number of false positives is still present after the image-based verification. Table 1 lists the number of initial hypotheses (those fulfilling the proximity criterion), the number of hypotheses surviving the refinement step and the number of final 3D points after image-based verification. The last column lists the execution times on a 2GHz AMD Athlon XP.



(a) (b) (c)

(d) (e) (f)

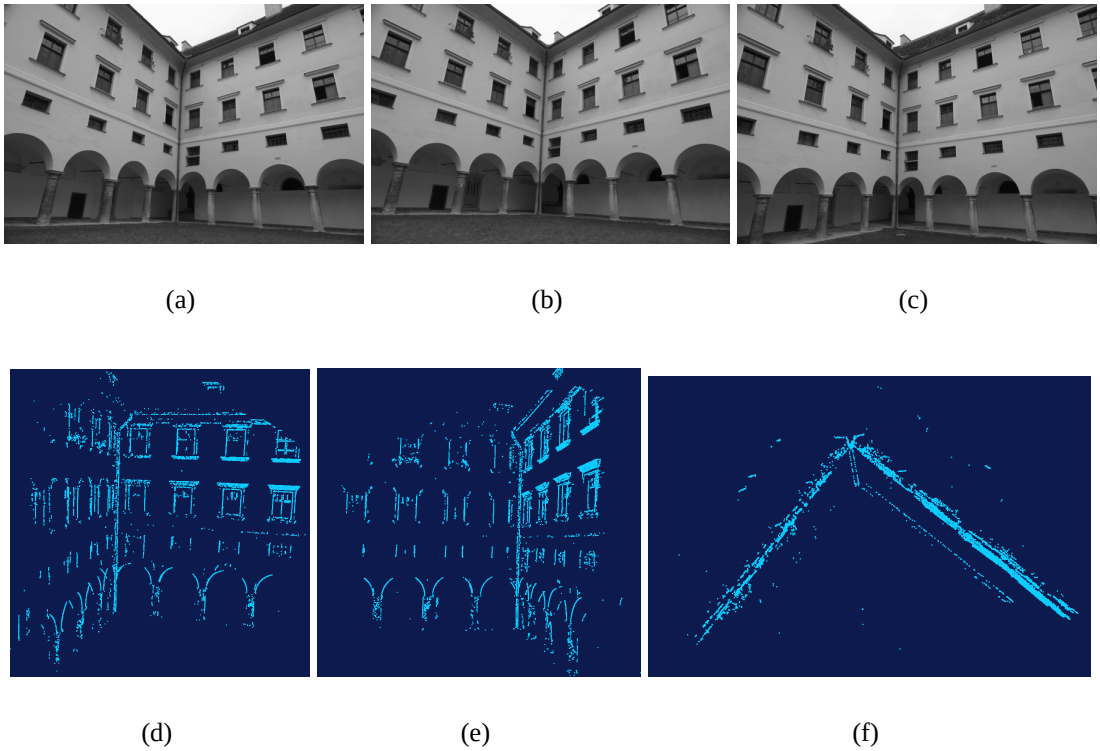
Figure 6: Top row: Three of the five images used of the synthetic turntable scene of a house (image size= 1000 $\times$ 750). Bottom row: Three views of resulting 3D point cloud.

4 Conclusions and Future Work

We presented a method for the fast computation of sparse 3D point sets from multiple oriented images. The main contribution is the introduction of distance transform for the accurate measurement of the re-projection error and the image-based outlier detection. This combination of feature-based constraints and an image-based similarity criterion allows a robust and efficient generation and verification of 3D hypotheses. Future work will include an analysis of the accuracy of the 3D hypotheses using synthetic data sets. We also want to investigate the influence of the size of the support region on the performance of the image-based similarity measure. So far we are only using gray-scale images, but the use of color images seems promising. The integration of color features could be implemented by expanding the orientation histogram to a more complex feature vector.

Acknowledgments

This work has been done in the VRVis research center, Graz and Vienna/Austria (http://www.vrvis.at), which is partly funded by the Austrian government research program Kplus. The authors also acknowledge the support from members of the Center for Machine Perception (CMP) in Prague, especially from Jana Kostkova and Radim Sara.



(a) (b) (c)

(d) (e) (f)

Figure 7: Top row: Three of the five images of the courtyard scene (image size = 2160 $\times$ 1440) Bottom row: Three views of the resulting 3D point cloud.

References

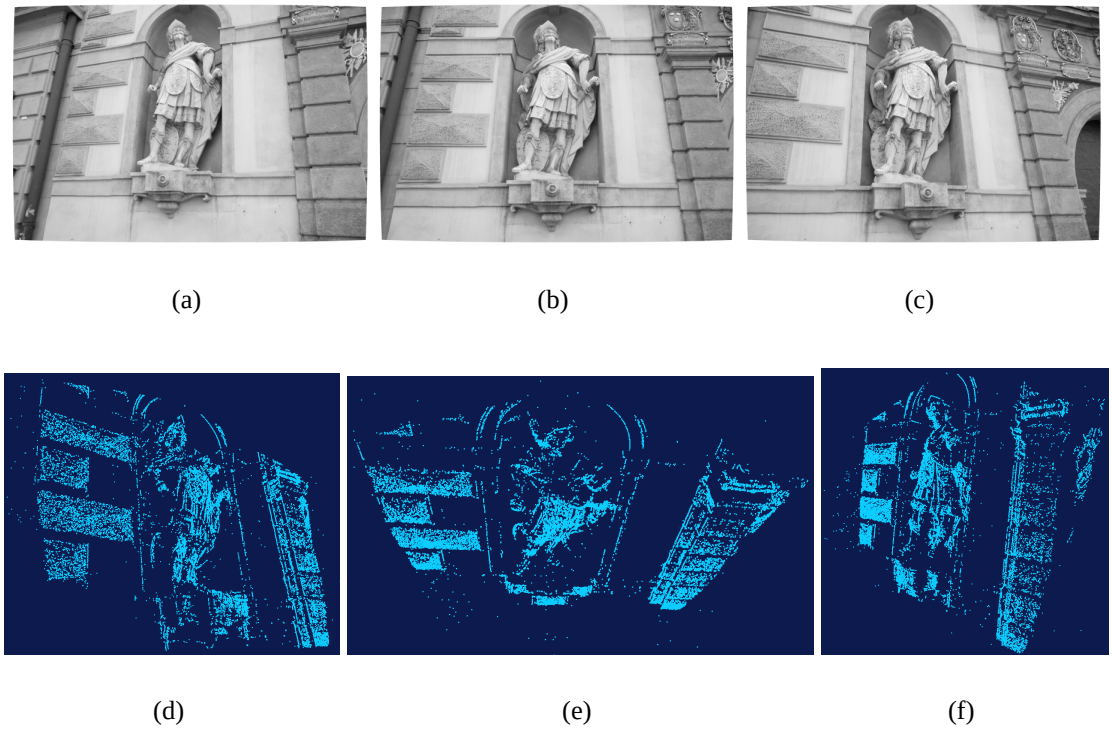
[1] G. Borgefors. Chamfering: A fast method for obtaining approximations of the Euclidean distance in N dimensions. In *Proc. 3rd Scand. Conf. on Image Analysis (SCIA3)*, pages 250–255, Copenhagen, Denmark, July 1983.

[2] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 8(6):679-698.*, 8(6):679–698, 1986.

[3] B. Huet and E. Hancock. Cartographic indexing into a database of remotely sensed images. In *WACV'96, pages 8–14, Dec 1996.*, 1996.

[4] Franck Jung, Vincent Tollu, and Nicolas Paparoditis. Extracting 3d edgels hypotheses from multiple calibrated images: A step towards the reconstruction of curved and straight object boundary lines. *ISPRS Journal of Photogrammetric Computer Vision*, B:100104, 2002.

[5] S. Kang, R. Szeliski, and J. Chai. Handling occlusions in dense multiview stereo. In *In IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, December 2001*, volume 1, pages 103–110, 2001.



(a) (b) (c)

(d) (e) (f)

Figure 8: Top row: Three of the five statue images (image size = 2032 $\times$ 1352). Bottom row: Three views of the resulting 3D point cloud.

[6] Kiriakos N. Kutulakos and Steven M. Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 38:199–218, 2000.

[7] David G. Lowe. Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision ICCV, Corfu*, pages 1150–1157, 1999.

[8] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Proceedings of the 1998 IEEE International Conference on Computer Vision, Bombay, India, January 1998*, pages 59–66, 1998.

[9] Steven M. Seitz and Charles R. Dyer. Photorealistic scene reconstruction by voxel coloring. In *Proc. Computer Vision and Pattern Recognition Conf.*, pages 1067–1073, 1997.

[10] Christoph Strecha, Tinne Tuytelaars, and Luc Van Gool. Dense matching of multiple wide-baseline views. In *ICCV03*, pages 1194–1201, Nice, France, 2003.