https://doi.org/10.2352/ISSN.2470-1173.2017.9.IRIACV-269

Line-Scan Stereo Using Binary Descriptor Matching and Regularization

Reinhold Huber-Mörk, Kristián Valentín, Bernhard Blaschitz, Svorad Štolc
Center for Vision, Automation and Control, AIT Austrian Institute of Technology, Vienna, Austria

Abstract

We present a line-scan stereo system and descriptor-based dense stereo matching for high-performance vision applications. Additionally we introduce a post-processing step based on total variation (TV) regularization for robust disparity estimation. Descriptor-based matching utilizes the Stochastic Binary Local Descriptor (STABLE). The performance of STABLE was shown to be superior to other binary descriptors, both w.r.t. stereo reconstruction quality as well as runtime performance. Regularized estimation of disparity maps is suggested as a hierarchical and iterative post-processing procedure where the Pseudo-Huber-TV norm was employed. We describe the hardware setup consisting of two line-scan cameras mounted in a car trailer and observing the road surface. Presented are results of 3D road surface reconstruction which are used in applications of road infrastructure maintenance.

Introduction

Depth information from images is typically obtained using active sensing, e.g. time-of-flight sensors [1], pattern projection [2], illumination variation by photometric stereo [3], or passive systems like focus variation [4], light field cameras [5] and multi-camera systems [6]. Line-scanning is a popular method to acquire images of moving objects, especially in machine vision applications. We utilize this acquisition principle, extended to binocular stereo, for an application in ground reconstruction from a vehicular platform. The application area is the inspection of road surface conditions.

Linear pushbroom cameras for earth-observating satellites are an established technology and have also been studied for stereo based ground reconstruction in various papers, e.g. by Gupta and Hartley [7]. Road surface reconstruction and free space estimation using area-scan cameras was reported by various authors, e.g. by Labayrade et al. [8]. Precise calibration of line-scan cameras, which is a prerequisite in stereo imaging, was discussed by Luna et. al [9] and by Caulier and Spinnler [10].

The STochAstic Binary Local dEscriptor (STABLE) for disparity estimation was discussed by Štolc et al. [11]. STABLE belongs to a broad class of local binary descriptors, along with the census transform (CT) [12], local binary patterns (LBP) [13], BRIEF [14] and others. Efficient representations and fast matching is obtained by the family of binary descriptors.

To incorporate prior knowledge, i.e. a smoothness assumption, and to cope with missing data we employed an cost minimization formulation where the cost from STABLE is forming our data term which is extended by adding a regularization term based on total variation [15]. An energy minimization procedure is iteratively solved by an optimization approach. The iterative scheme is embedded in a hierarchical representation. This enhances details significantly when compared to a purely data-dependent solution.

This paper is organized as follows. First we will present the line-scan stereo setup and descriptor based matching, then we present our hierarchical regularization method for cost stacks derived from stereo matching. Results for synthetic and real data are presented. Finally, we draw conclusions.

Line-scan stereo

We describe the acquisition setup and calculation of the stereo matching costs using the suggested STABLE descriptor. These matching costs will further be used in the regularization framework.

Line-scan stereo image acquisition

In stereo imaging the range for each pixel is obtained from the estimated disparity, i.e. the displacement between corresponding points observed in two (or more) images. The epipolar constraint in a binocular stereo vision system states that a point in one image is found along the corresponding epipolar line in the other image. Epipolar rectification for area-scan stereo pairs aligns epipolar lines to images lines, thus reducing the correspondence estimation by searching an expected disparity range oriented along image lines. In a line-scan stereo system one mechanically adjusts this geometrical constraints such that epipolar lines correspond to sensor lines. Estimation of disparities is then performed along sensor lines, accordingly.

One or two line-scan stereo cameras which are sensitive in the visible spectrum for acquisitions of the ground surface while the acquisition device is moving could be used for the considered purpose. In line-scan one can either use one long image sensor line shared by two lenses or two collinearly arranged line-scan image sensors observing the same surface line patch. Our setup uses two collinearly arranged line-scan sensors observing the surface from two different viewpoints. The optical axes are verged in order to obtain a larger overlapping region, see Fig. 1 a) and b). Calibration ensures the alignment of the sensors such that the plane spanned by the left optical axis and left sensor line is coplanar with the plane spanned by the right optical axis and right sensor line. This property is important to fulfill the epipolar constraint at each depth and requires a calibration procedure which ensures collinearity of the sensor lines at a number of distances [16].

Verging of the optical axis has two drawbacks. First, the object resolution decreases from left to right in one view and from right to left in the other view, respectively. Secondly, the limited depth of field might result in sharpness reduction depending on

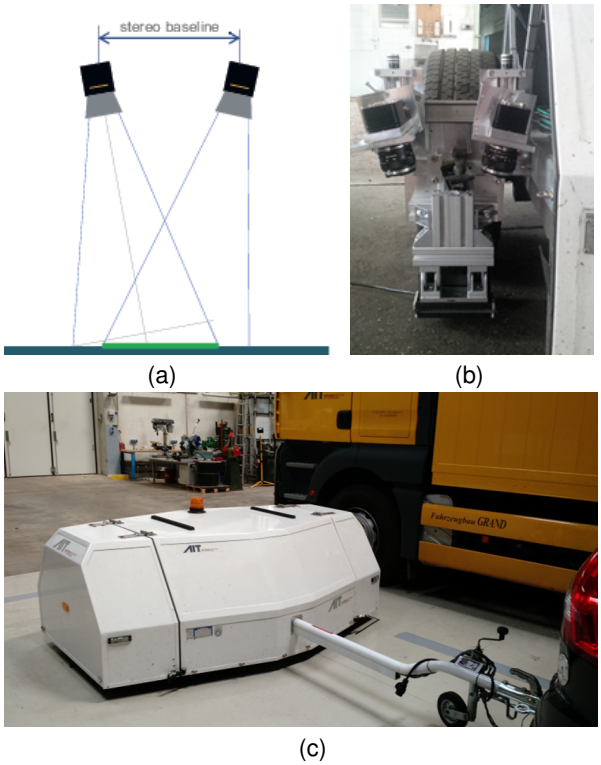

Figure 1. *Binocular stereo image acquisition: (a) sketch of verged stereo geometry for two collinearly arranged line-scan sensors,(b) cameras mounted in trailer, (c) car trailer carrying the imaging devices.*

optical parameters and adjustment when compared to a canonical stereo system. Geometric calibration of the sensor lines ensures a constant object pixel size at the regular working distance for planar surfaces.

The depth of field was estimated to be on the order of magnitude of ± 6.16mm for a f-number of 5.6, a magnification of 0.1 and a sensor pixel size of 10μm. For f-numbers of 1.4 or 2.8 we would obtain a depth of field of ± 1.54mm or ± 3.08mm, respectively. Although these are quite low numbers it turned out to be sufficient to compensate for the varying distance due to verging and the expected depth variation in road inspection. Other parameters of the system are a baseline of 220mm, a working distance of 480mm and a verging of the obigal axes of 11° w.r.t. the surface normal.

Descriptor based stereo matching

The STABLE descriptor is a binary descriptor, where pixel pairs at random positions inside a matching window are compared and mapped to descriptor bits. With STABLE we are able to map a larger number of pixel pairs to a smaller number of descriptor bits. The most similar descriptor is BRIEF. The BRIEF descriptor uses a (typically sparse) sub-sample of pixel pairs located at arbitrary positions in the matching window. The resulting descriptor lengths equals the number of pixel pair comparisons performed. With STABLE we also get pixel pairs at random positions, but we are able to map a larger number of pixel pairs to a smaller number of descriptor bits.

The STABLE descriptor considers an image patch $\mathbf{p}$ of size $X \times Y$ pixels. The operation β derives the i-th descriptor bit $d_i \in \mathbf{d}$ from patch $\mathbf{p}$ as follows:

$$\beta(\mathbf{p},i) = \begin{cases} 1 & \text{if } (\mathbf{p} * f_i) > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where f_i is a filter mask of equal size as the image patch $\mathbf{p}$. We refer to the operation β as the *binarized convolution*. The filter dictionary $\mathbf{f}$ contains K sparse filter masks f_i. Each entry in f_i is either 0, 1 or -1. The descriptor $\mathbf{d}$ is a K-dimensional bitmask which is obtained for a given image patch $\mathbf{p}$ using

$$\mathbf{d}(\mathbf{p}) = \sum_{i=1}^{K} 2^{i-1} \beta(\mathbf{p},i). \tag{2}$$

As with other binary descriptors, the Hamming distance is used for matching descriptor pairs.

We consider a discrete range of disparities $[r_1,\ldots,r_P]$ for which the descriptors corresponding to each image pixel position are compared by the Hamming distance. The result is a cost stack C of dimension $M \times N \times P$, where M and N are the image dimensions and P is the number of evaluated disparities.

Regularization

Regularization of the cost stack C is suggested as a hierarchical and iterative procedure. A regularized solution S is sequentially updated using local neighborhood information, i.e. horizontally and vertically neighboring pixels. The regularization is penalizing large disparity variations in local neighborhoods. This property is propagated through iteration and embedded in a hierarchical framework. The algorithm starts with the cost stack $C_{i_{max}}$ from STABLE matching at the coarsest pyramid level i_{max}. An initial solution for the coarsest pyramid level is given by the minimum of the cost stack at each pixel position (x,y) for the disparities $z = 1 \ldots P$ as follows:

$$S^1_{i_{max}}(x,y) = \underset{z=1\ldots P}{\arg\min} C_{i_{max}}(x,y,z). \tag{3}$$

The solution at the i-th pyramid level and j-th iteration is then iteratively refined:

$$S_i^{j+1}(x,y) = \begin{cases} \underset{z=1\ldots P}{\arg\min} \left(C_i(x,y,z) + \lambda L^j(x,y,z)\right) & \forall (x,y) \in \Omega^{j+1}, \\ S_i^j(x,y) & \text{otherwise,} \end{cases} \tag{4}$$

where $L(\cdot)$ is the regularizer function, λ controls the regularization strength, and Ω^{j+1} is the set of pixel coordinates to be updated in this iteration. Finally, $S_1^{j_{max}}$ is accepted as the regularized solution at the original resolution.

Unlike common procedures in regularization, where the direction of optimization, i.e. along lines or along columns, is changed between adjacent iteration steps we suggest updating pixels in two complementary groups in each step. In order to ensure convergence we suggest a checkerboard-like update pattern which switches between adjacent iterations. The set of pixel coordinates updated in j-th iteration is defined as follows:

$$\Omega^j = \{(x,y) \mid \text{if } x+y+j \text{ is even}\}. \tag{5}$$

The Pseudo-Huber loss [17] is used to regularize the information originating from matching costs at each pixel position and its neighborhood:

$$L^j(x,y,z) = \sum_{(\hat{x},\hat{y})\in\mathcal{N}_4(x,y)} \delta^2\left(\sqrt{1+|S^j(\hat{x},\hat{y})-z|^2/\delta^2}-1\right), \quad (6)$$

where $(\hat{x},\hat{y})$ are pixel positions within the 4-neighborhood $\mathcal{N}_4(x,y)$. The Pseudo-Huber loss function has the nice property of having linear (i.e. total variance) behavior towards large values, while being quadratic around zero. The parameter δ governs the range of the quadratic behavior around zero.

Finally, the regularized cost stack at the original resolution S_1 is searched for the minimum cost at each pixel which provides the associated disparity estimation. Fig. 2 outlines the suggested algorithm. Intermediate results are shown in Fig. 3 where results for a pyramid of eight levels is constructed and the disparity is iteratively refined at each level. The solution at each level is upsampled and used as an initial solution for refinement at the next finer level.

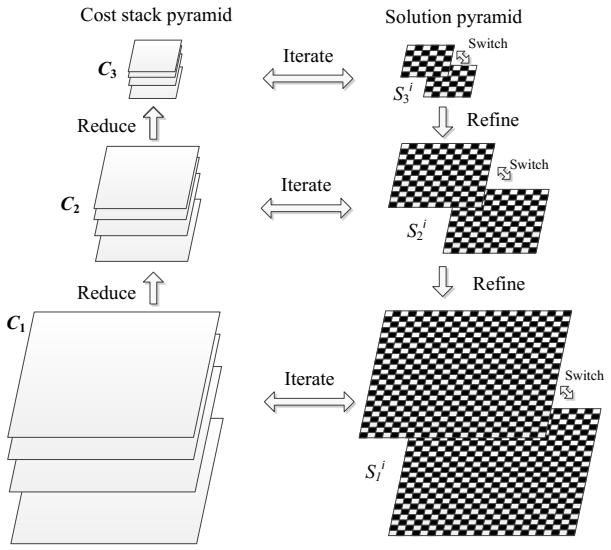

Figure 2. *Regularization framework: the initial cost stack C_1 is reduced in an image pyramid (in this case in two steps). The initial solution S_3^1 is set to the minimum cost of C_3 and iteratively regularized using an alternating checkerboard-pattern. The solution is then upsampled to the finer resolution and again iteratively regularized, and so on. The final result is obtained at the original resolution after regularization.*

Results

We will present the depth reconstruction performance for synthetic data with ground truth as well for real-world data of a road surface.

Synthetic Data

We used 3D surface models to evaluate our algorithm in a quantitative way. The models consist of real-world surfaces that were captured with a camera and a calculated depth map [18]. These scenes were rendered using POV-Ray [19], with a virtual construction similar to our real-world setup. Two cameras were set up to look under the angle towards a central point on a surface located at a defined distance.

Fig. 4 shows the surface images, ground truth depth maps, our results and error maps. Error rates were measured in disparity values by the mean squared error (MSE), the mean absolute error (MAE), bad1, which represents the number of pixels with an error $err > 1$ of the disparity estimation result compared to the total amount of pixels, as well as error MAE, which is the mean absolute value divided by disparity range. The surfaces cover disparity ranges from 9.06 (gravel) to 10.17 (cobble). The results shown in Table 1 were achieved by a 64-bit STABLE descriptor operating on 25×25 pixel windows refined by the suggested regularizer.

Quantitative reconstruction results for synthetic surfaces (MAE, MSE are measured in disparity units).

	earth	cobble	gravel
disparity range	9.93	10.17	9.06
MSE	0.35	0.17	0.29
MAE	0.46	0.29	0.43
bad1 (%)	5.13	2.02	7.07
error MAE (%)	4.67	2.86	4.74

Road Data

Due to the lack of ground truth we refer to a manual annotation of interesting properties which are visible to human observers and show the derived 3D reconstruction from which these properties become clearly visible. In most of the results there is a vertically oriented 3D structure visible. This stems from diamond grinding, which is a pavement preservation technique to remove surface irregularities in order to reduce noise and increase safety.

Fig. 5 a) shows an image of a ground concrete road surface with an expansion joint. The grinding stripes, as well as the expansion joint, are visible in the disparity map in Fig. 5 b). The brighter the disparity the closer the observed object point is to the observer. Figs. 5 c) and d) show a ground concrete pavement with a small hole. A larger break out of the surface is shown in Figs. 5 e) and f). Finally two grinding lanes of different depth are provided in Figs. 5 g) and h). Additionally, in the upper left corner there is material, presumably a chewing gum, observed in the area of the deeper grinding.

Conclusion

We have presented the hardware details and algorithms for a line-scan stereo system for close-range surface observation from a mobile platform. Illustrative qualitative results are provided on real-world data and a quantitative evaluation of our approach was shown on synthetic data with ground truth information. Quantitative results showed an average MAE smaller than 5% in relation to the disparity range of the ground truth surface. The visual evaluation additionally shows that the results are well suited for the task of road surface assessment, especially when compared to the state-of-the-art procedures. Further work will include stabilization of the setup and speedup of the computation, especially the regularization step.

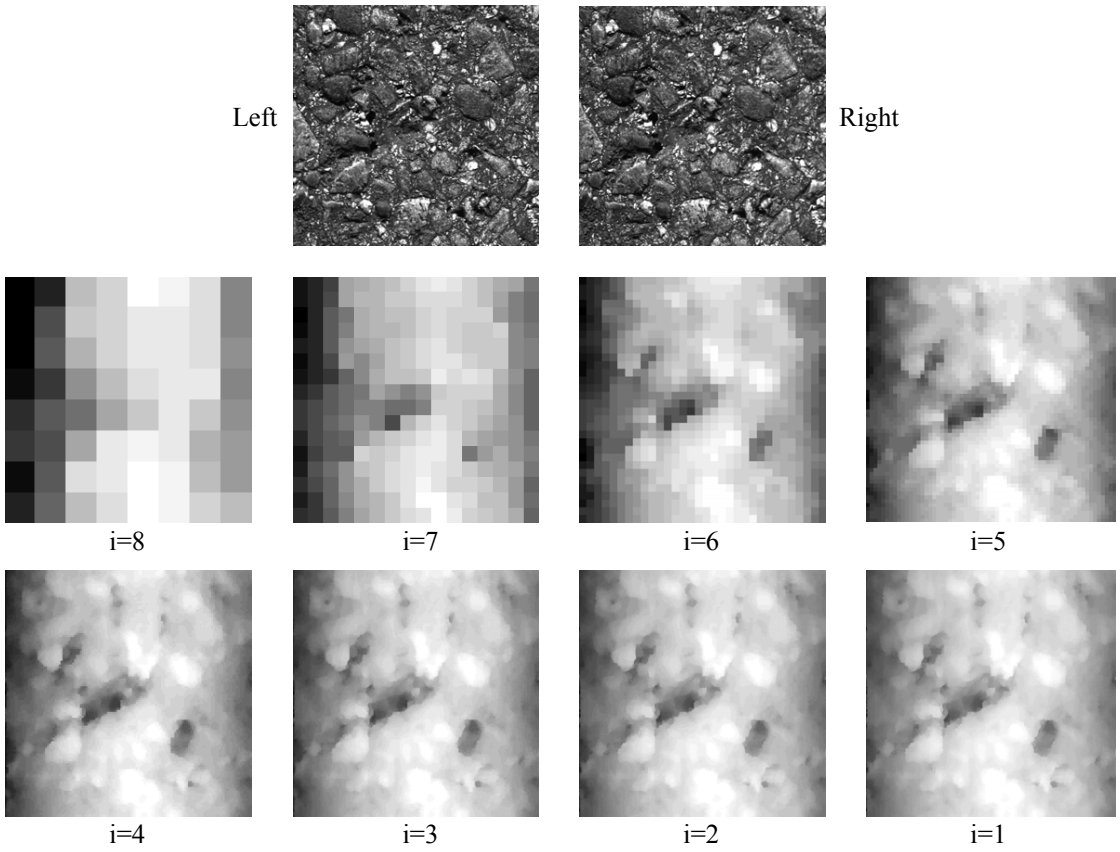

Figure 3. *Hierarchical refinement of the disparity solution: The stereo pair is shown on top. The sequence of iteratively refined solutions for the disparity for pyramid levels starting from the coarsest resolution $i = 8$ to the original resolution $i = 1$ are shown below (for viewing purposes scaled to the original solution).*

References

[1] S.B. Gokturk, H. Yalcin, and C. Bamji. A time-of-flight depth sensor - system description, issues and solutions. In *Proc. of Computer Vision and Pattern Recognition Workshop*, June 2004.

[2] Z. Zhang. Microsoft Kinect sensor and its effect. *IEEE MultiMedia*, 19(2):4–12, April 20012.

[3] R. Basri, D. Jacobs, and I. Kemelmacher. Photometric stereo with general, unknown lighting. *Int. J. of Computer Vision*, 72(3):239–257, 2007.

[4] E. Krotkov and Martin J.P. Range from focus. In *Proc. of Intl. Conf. on Robotics and Automation*, pages 1093–1098, 1986.

[5] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan. Light field photography with a hand-held plenoptic camera. Technical Report CSTR 2005-02, Stanford Univ., April 2005.

[6] M. Okutomi and T. Kanade. A multiple baseline stereo system. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(4):353–363, 1993.

[7] R. Gupta and R. Hartley. Linear pushbroom cameras. *IEEE Trans. on pattern analysis and machine intelligence*, 19(9):963–975, 1997.

[8] R. Labayrade, D. Aubert, and J.-P. Tarel. Real time obstacle detection in stereovision on non flat road geometry through v-disparity representation. In *Proc. of Intell. Vehicle Symp.*, pages 646–651, 2002.

[9] C. A. Luna, M. Mazo, J. L. Lazaro, and J. F. Vazquez. Calibration of line-scan cameras. *IEEE Transactions on Instrumentation and Measurement*, 59(8):2185–2190, Aug 2010.

[10] Y. Caulier and K. Spinnler. Calibration of 1D cameras determination of 3D reconstruction accuracy. In *Proc. of Vision, Modeling and Visualization (VMV)*, pages 55–62, Stanford, CA, USA, November 2004.

[11] S. Štolc, K. Valentín, and R. Huber-Mörk. STABLE: Stochastic binary local descriptor for high performance dense stereo matching. In *Proc. of IS&T Intl. Symp. on Electronic Imaging: Machine Vision Applications IX*, February 2016.

[12] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In *Proc. of Europ. Conf. on Computer Vision (ECCV)*, pages 151–158, Stockholm, SE, 1994.

[13] T. Mäenpää. *The local binary pattern approach to texture analysis - extensions and applications*. PhD thesis, Machine Vision and Media Processing Unit, Infotech Oulu, Univ. of Oulu, Finland, 2003.

[14] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief: Binary robust independent elementary features. In *Proc. of Europ. Conf. on Computer Vision (ECCV)*, pages 778–792, 2010.

[15] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

[16] D. Antensteiner, B. Blaschitz, C. Eisserer, R. Huber-Mörk, J. Ruisz, S. Štolc, and K. Valentín. Line-scan stereo for 3D ground reconstruction. In *Proc. of KIT/Fraunhofer IOSB Forum Bildverarbeitung*, Karlsruhe, D, December 2016.

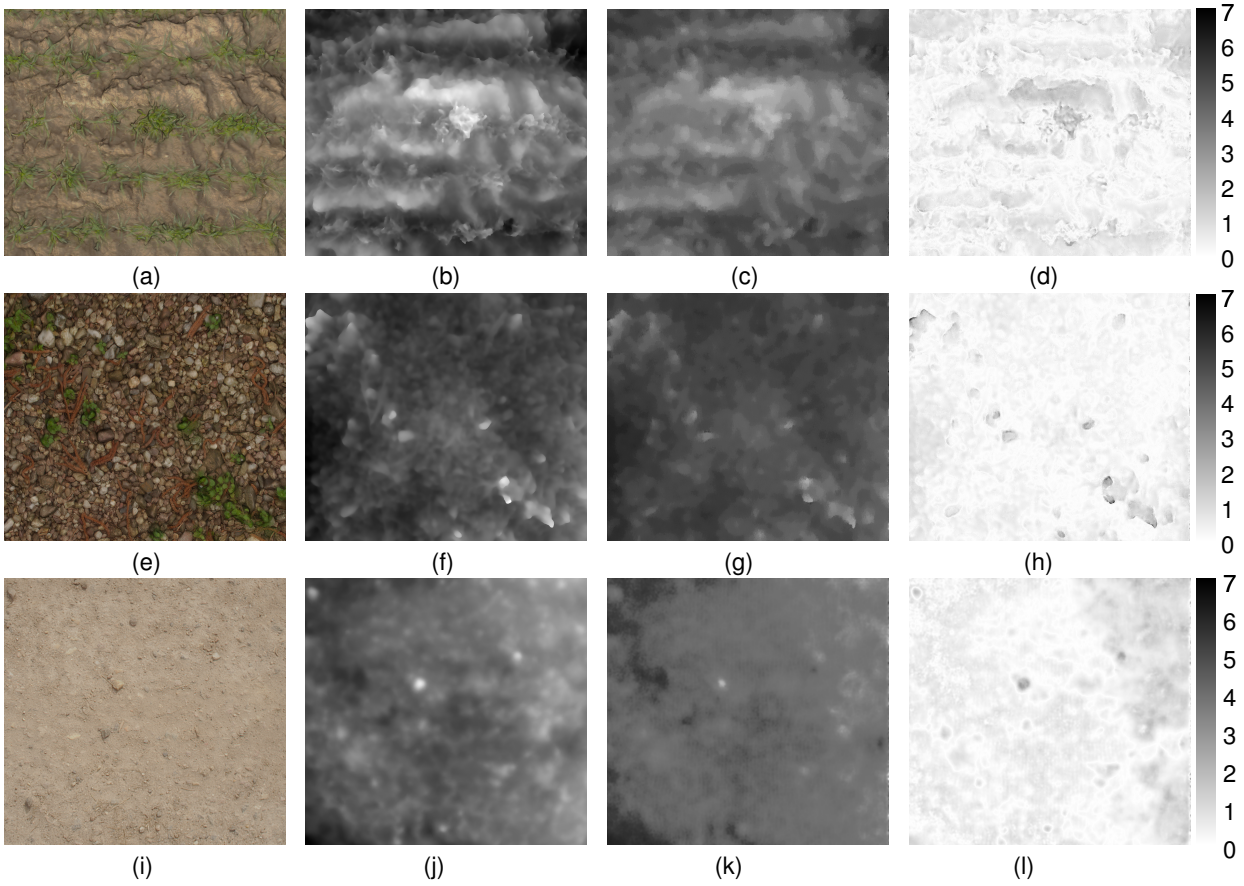

Figure 4. *Results for synthetic data: grass on earth surface: (a) image, (b) ground truth, (c) our result (d) error map; cobble stones: (e) image, (f) ground truth, (g) our result (h); gravel stones: (i) image, (j) ground truth, (k) our result (l) error maps. In the error maps, which are the absolute differences between ground truth and disparities, white indicates no depth error, black indicates large errors (scale from 0 to 7).*

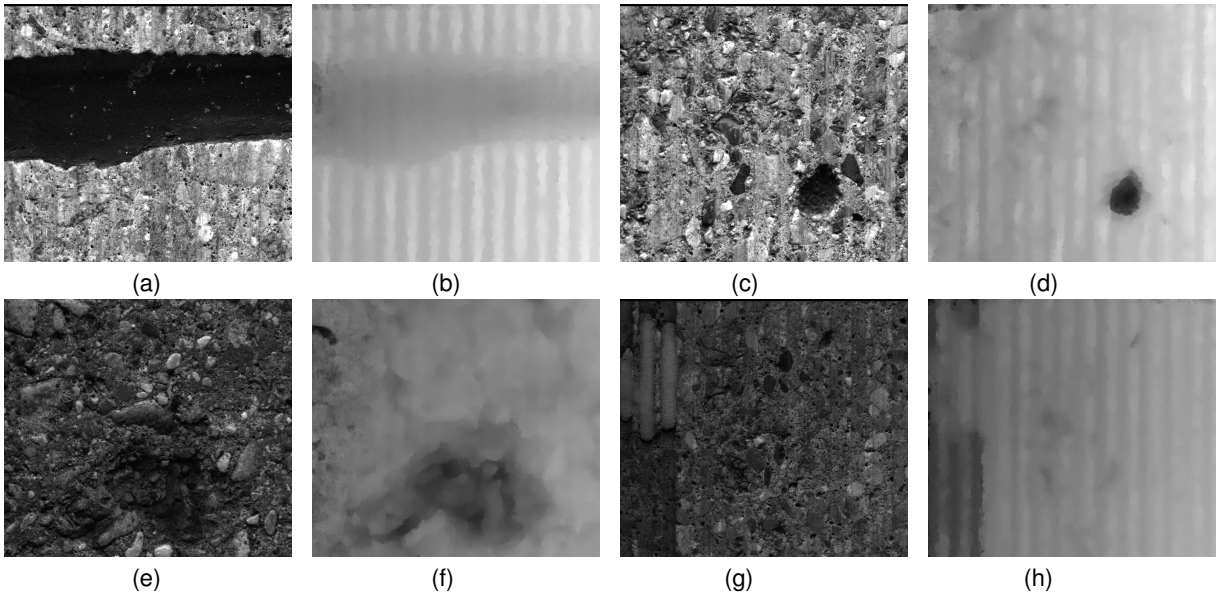

Figure 5. *Illustrative results for road surface data (disparity maps smoothed by TV regularization): concrete surface with expansion joint after grinding (a) image, (b) disparity; concrete surface with a hole after grinding (c) image, (d) disparity; concrete surface with a larger break out region (e) image, (f) disparity; a chewing gum like object observed after grinding at different depths (g) image, (h) disparity.*

[17] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Transactions on Image Processing*, 6(2):298–311, February.

[18] Siebencorgie. Scanned surface collection. `http://siebencorgie.jimdo.com/`, 2016. [Online; accessed 22-Sept-2016].

[19] Persistence of Vision Pty. Ltd. Scanned surface collection. `http://www.povray.org/download/`, 2016. [Online; accessed 22-Sept-2016].

Author Biography

Reinhold Huber-Mörk received his PhD in computer science from the University of Salzburg, Austria, in 1999. Since then he worked at the Aerosensing GmbH, Oberpfaffenhofen, Germany, in remote sensing, at the Advanced Computer Vision GmbH, Vienna, Austria, in computer vision and in 2006 he joined the AIT, Vienna, Austria, where he is currently senior scientist in the field of machine vision.

Kristián Valentín received his PhD in Computer Science from Comenius University, Bratislava in 2015. Since 2014, he worked at AIT, Vienna in the field of computational imaging and computer vision. In 2017, he joined Photoneo, Bratislava to work on 3D scanners.

Bernhard Blaschitz earned his masters degree in Mathematics from the University of Vienna in 2008 and a PhD degree in Applied Geometry from Technical University of Vienna, Austria in 2014. He joined the AIT Austrian Institute of Technology in 2015 and works as a researcher at the Center for Vision, Automation & Control.

Svorad Štolc is a scientist AIT, Vienna. In 2002, he earned his masters degree in Computer Science from Comenius University, Bratislava and, in 2009, PhD degree in Bionics and Biomechanics from Technical University, Košice and Slovak Academy of Sciences, Bratislava. He is a (co)author of more than 50 peer-reviewed scientific papers and holds a number of patents in machine vision. His main research areas are computational imaging and machine vision with a focus on industrial inspection and document security.