Evaluating the Suitability of Feature Detectors for Automatic Image Orientation Systems

Timo Dickscheid and Wolfgang Förstner

Department of Photogrammetry
Institute of Geodesy and Geoinformation
University of Bonn
dickscheid@uni-bonn.de, wf@ipb.uni-bonn.de

Abstract. We investigate the suitability of different local feature detectors for the task of automatic image orientation under different scene texturings. Building on an existing system for image orientation, we vary the applied operators while keeping the strategy fixed, and evaluate the results. An emphasis is put on the effect of combining detectors for calibrating difficult datasets. Besides some of the most popular scale and affine invariant detectors available, we include two recently proposed operators in the setup: A scale invariant junction detector and a scale invariant detector based on the local entropy of image patches. After describing the system, we present a detailed performance analysis of the different operators on a number of image datasets. We both analyze ground-truth-deviations and results of a final bundle adjustment, including observations, 3D object points and camera poses. The paper concludes with hints on the suitability of the different combinations of detectors, and an assessment of the potential of such automatic orientation procedures.

1 Introduction

1.1 Motivation

Automatic image orientation has become mature even in close-range and wide-baseline scenarios with significant perspective distortions between overlapping views. Fully automatic solutions of the relative orientation problem are available for such cases, relying on rotation and scale invariant [1] or even fully affine invariant correspondence detection techniques [2–4]. Such systems however do not always perform well: It will turn out that the suitability of detectors varies especially depending on the 3D structure and texturedness of the surfaces.

The applications of automatic image orientation are manifold. Examples are the alignment overlapping subsets of unordered image collections, known as the "stitching problem", which requires to recover the relative positioning of the cameras [5], or the automatic computation of 3D scene models from images, where one needs accurate estimates of the extrinsics for computing dense 3D point clouds.

For evaluating variations of an automatic image orientation system, one may consider two cases: (i) Given a fixed strategy, what is the impact of different

Fig. 1. Some example images of the datasets used. Top left: ENTRY-P10, top right: HERZ-JESU-P25, bottom left: EMPTY-2, bottom right: GLCUBE-TEXTURE/GLCUBE-COAST.

operators on the result? (ii) Given a specific operator, how successful are different strategies in solving the problem? In this contribution, we are concerned with (i) and leave the orientation strategy fixed.

We will continue by giving a short overview on the state of the art in local feature detection and automatic image orientation, before describing the system used for this evaluation in section 2, together with the applied keypoint detectors. The experimental setup is detailed in section 3, followed by an analysis of the results in 4. We conclude with a short summary and outlook in section 5.

1.2 Related Work

Fully automated systems for solving the relative orientation problem are available since several years [6–8]. The procedure used in our experiments is based on [9], which uses Lowe features [1] for automatic correspondence detection, and related to the approach of [10]. There is a lot of recent work on optimizing such procedures. To only mention a few, in [11] it was shown how to connect pairwise epipolar geometries by first registering the rotations and then globally optimizing the translation vectors in a robust manner, while the authors of [12] use a small Bayesian network for pairwise correspondences in image triples in order to make efficient and statistically sound use of the correspondence information.

Several good feature detectors have been established in the last years. Beyond the classical junction detectors [13, 14], based on the second moment matrix computed from the squared gradients, the influential work of Lowe [1] showed that robust automatic correspondence detection is possible under significant illumination and viewpoint changes. Lowe uses a detector searching for local maxima of the Laplacian scale space, yielding scale invariant dark and bright blobs, and computes highly distinctive yet robust "SIFT" descriptors for these local image patches. Subsequent developments brought detectors with invariance under affine distortions for blobs and regions [2–4], which is favorable under very strong viewpoint changes. Recently, a robust scale-invariant junction detector has also

been proposed [15]. All these detectors can just as well exploit the power of SIFT descriptors for automatic correspondence analysis.

2 A System for Automatic Image Orientation

2.1 Image Orientation Strategy

We follow the scheme published in [9] which will be shortly summarized here. As an input, we assume an unsorted set of N overlapping images with known intrinsics, along with a set of K_n local features each, i.e. $F_{nk} = \{x_{nk}, y_{nk}, \theta_{nk}, \mathbf{d}_{nk}\}$ with $0 < n \leq N$ and $0 < k < K_i$. Here, (x_{nk}, y_{nk}) is the location of the k-th feature in the domain of image n, and θ_{nk} is an additional geometric description of the feature window, possibly its scale σ_{nk} or a matrix A_{nk} containing complete ellipse parameters. The $1 \times M$-vector $\mathbf{d}_{nk}$ is a distinctive description of the feature, in our case a SIFT descriptor [1] with $M = 128$. It is computed over the local neighborhood coded by θ_{nk}.

The procedure starts by comparing descriptors of all possible image pairs (n, m), yielding sets $C_{nm} = \{(p, q) \mid 0 < p < K_n,\ 0 < q < K_m\}$ of initial correspondences. As the intrinsics are assumed to be known, we compute the relative orientation of each image pair using Nister's 5-Point algorithm [16] embedded into a RANSAC scheme [17, 18]. This computation step not only allows for robust approximate values for the pairwise epipolar geometries (denoted as *EG's* in the following), but also acts as a filter on the initial correspondences, usually yielding updated sets C_{nm} with significantly reduced outlier rates.

Based on the filtered sets C_{nm}, we can now directly determine multiview correspondences from the pairs through simple index propagation. The EG's for image pairs are then connected in an iterative manner, prioritized by their quality, which is based on the number of valid coplanarity constraints. Note that some 3-fold correspondences are required to determine the scale between connected EG's. The system only yields one set of connected image orientations: In case that no further EG can be connected, the procedure stops, regardless of another isolated EG cluster.

Subsequently, triplet tests are carried out for further elimination of invalid EG's: For each triple of connected orientations, the product of their rotation matrices has to equal the identity matrix, and the baselines have to be coplanar. After determining 3D object points from the multiview correspondences, the whole block is optimized by a sparse bundle adjustment [19].

2.2 Applied Feature Detectors

Other than proposed in [9], we try different input feature sets F_{nk}. This is motivated by the fact that the Lowe detector (denoted as LOWE in the following) alone is not always the best choice, though most often a good one. Consider the image pair in the bottom left of Fig. 1: The amount of texture is critically low here, and it will turn out that the system is not able to successfully process

the whole dataset using only LOWE. We will therefore also present experimental results obtained when using the popular Harris and Hessian affine detectors [2], denoted by HARAF and HESAF, and the Maximally Stable Extremal Regions detector [3, MSER]. Furthermore, we use a scale-invariant junction detector as recently proposed in [15]. Note that the junction features are only a subset of the detector output, determined by restricting to $\alpha = 0$ in [15, eq. (7)]. Lastly we include a new detector based on information theory, which will be described shortly in the following.

*Maximum-entropy-detector (*ENTROPY*).* The maximum entropy detector has been proposed in [20]. It is motivated by the idea of good image coding: We search for local patches of varying size, centered at each pixel position, with locally maximal entropy. Therefore at each position $(\mathbf{x}, \sigma, \tau)$ in the 3D scale space obtained by fixing $\sigma = 3\tau$, we compute

$$H(\mathbf{x}, \sigma, \tau) = k\sqrt{\frac{\lambda_2(M; \tau, \sigma)}{V_{\mathbf{x}}(\sigma)}} \log^2\left(\frac{V_{\mathbf{x}}(\sigma)}{V_n(\tau)}\right) \tag{1}$$

Here, $V_{\mathbf{x}}$ denotes the variance of the image intensities within the patch, which can be determined by averaging finite differences of the grayvalues, and V_n is the noise variance at the respective scale level, which can be analytically determined from a given noise estimate of the original image. The result is up to an unknown factor k, which does not affect the maximum search.

3 Experiments

Image Data. We report results for six image datasets, providing a range of different texturings and surface types:

1. The ENTRY-P10- and HERZ-JESU-P25-datasets provided by the authors of [10], at reduced resolutions of 512×768 and 256×384 [pel], respectively (see top row of Fig. 1). The datasets are provided with ground-truth projection matrices. We included HERZ-JESU-P25 especially for having a dataset with full 3D structure, following [21] who pointed out that this is a critical aspect of detector evaluations.
2. Our own EMPTY-1- and EMPTY-2-datasets with a resolution of 512×768 [pel] showing indoor scenes with very low amount of texture (bottom left of Fig. 1). These especially difficult datasets are a challenge for state-of-the-art orientation procedures.
3. Two artificial datasets GLCUBE-TEXTURE and GLCUBE-COAST resulting from a 3D graphics simulation of a cube observed from inside, with natural images as wallpaper textures, rendered at a resolution of 600×800 [pel]. For the texturing we have chosen samples of well-known image datasets from texture analysis and scene category recognition [22, 23]. One example pair of each set is shown on the bottom right of Fig. 1.

Investigated feature sets. We computed results (i) for each of the detectors individually, (ii) for all possible pairs complementing LOWE and SFOP, and (iii) for some promising combinations of three or four detectors. The settings of the orientations procedure were otherwise kept constant. The focus on LOWE and SFOP among the pairwise combinations is chosen due to the limited space in the paper, considering that LOWE and SFOP have shown to be most successful.

Indicators. After automatically computing the relative orientation of the images with the system described in section 2.1, we analyzed the following key indicators for each of the combinations and datasets:

1. The percentage P_O of successfully oriented images w. r. t. the overall number of images in a dataset, indicating success in calibrating the whole dataset.
2. The average standard deviation of observations $\widehat{\sigma}_{x'}$ as estimated by the bundle adjustment, reflecting the accuracy of observations.
3. The average number $\overline{N}_I$ of 3D object points observed in an image, indicating the stability of the estimated orientation for each particular image.
4. The average number $\overline{N}_O$ of independent observations of the 3D object points in overlapping images, indicating stability of the estimated camera poses.
5. The ratio C between the convex hull of observations and the image plane, as an indicator for good coverage of the image with observations.
6. The average deviation $\overline{D}_{X_0}$ of the estimated projection centers from the ground truth, where available, giving insight into the quality of the estimation.

Note that the differences $\overline{D}_{X_0}$ are computed after a coordinate transformation of the estimated frames into the coordinate system of the ground truth data, using the least squares solution proposed in [24].

As the results vary due to the RANSAC component of the system, we show average values over ten repeated estimates throughout the paper, along with the corresponding standard deviations depicted by black markers.

4 Results

Overall Suitability for Image Orientation. From Fig. 2 we see that not all datasets were successfully calibrated using separate detectors. Only the SFOP detector seems to handle all considered situations. ENTROPY at least solved the problem for all but EMPTY-2. The LOWE and MSER detectors work well with good and medium amount of texture, but yield incomplete results on the difficult EMPTY-2 and EMPTY-1 datasets. Both the HARAF and HESAF detectors yield incomplete results in all cases. Using combinations of two or three detectors however, we were usually able to get complete estimates. Only for EMPTY-2 and EMPTY-1, either LOWE, SFOP or ENTROPY were required for a successful combination.

Using the combination of LOWE and HESAF on GLCUBE-COAST, only 80% of the cameras were calibrated on average, although LOWE alone worked well.

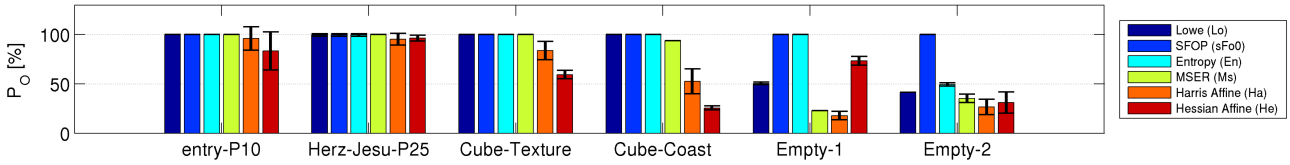

Fig. 2. Percentage P_O of successfully oriented cameras w. r. t. the overall number of cameras per dataset for individual detectors. Throughout the paper, the coloured bars show the mean over 10 repeated estimates, while the black bars denote the standard deviation.

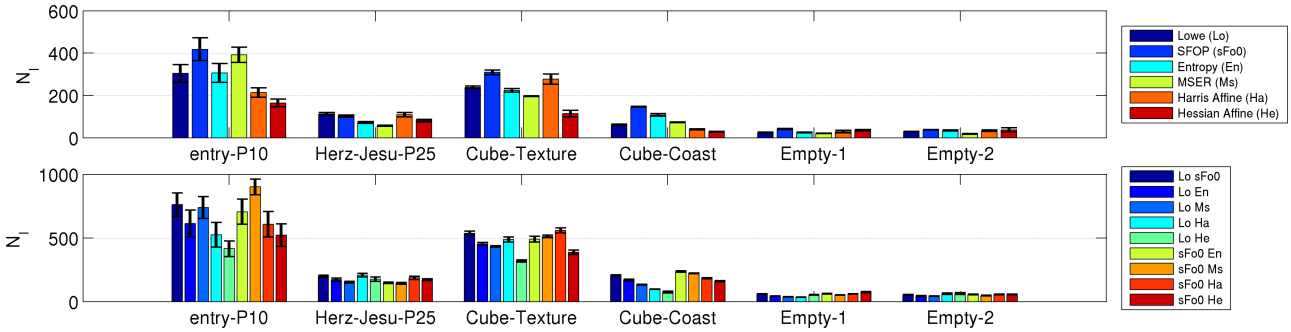

Fig. 3. Average number $\overline{N}_I$ of observed object points per image for individual detectors (top) and pairwise combinations (bottom).

Such negative interaction between two detectors is otherwise rarely observed. We believe that this is due to the fact that both detectors are based on the Laplacian, thus having highly redundant feature sets. One might hence conclude that combinations of very similar detectors should be avoided.

Repeatability and Amount of Observations. The average number $\overline{N}_I$ of object points observed in an image is often highest for SFOP among individual detectors (Fig. 3), while usually some of the other detectors yield comparable scores on particular datasets. On HERZ-JESU-P25 however, LOWE proves best. HESAF has the lowest score in many datasets. In case of pairwise combinations, the $\overline{N}_I$ approximately add as expected. The average number $\overline{N}_O$ of independent observations is significantly better for SFOP junction points on EMPTY-2 and EMPTY-1 (Fig. 4). It is an indicator for the repeatability, and underlines the importance of junction points for processing images of such scenes with poor texture. For combinations of detectors, we get mostly threefold points on average.

Average Accuracy of Observations. The average estimated standard deviation $\widehat{\sigma}_{x'}$ of the observations is worse for ENTROPY and HESAF compared to that of other detectors (Fig. 5). For ENTROPY this may be caused by the lack of a

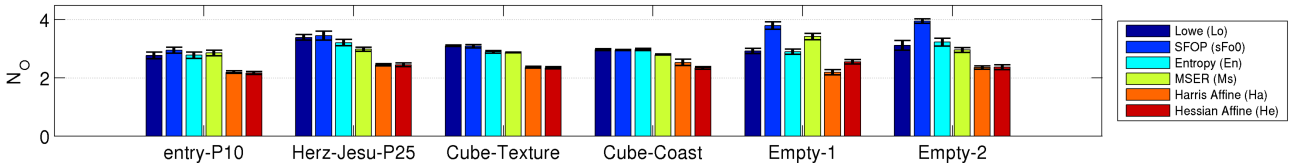

Fig. 4. Average number $\overline{N}_O$ of independent observations of 3D object points.

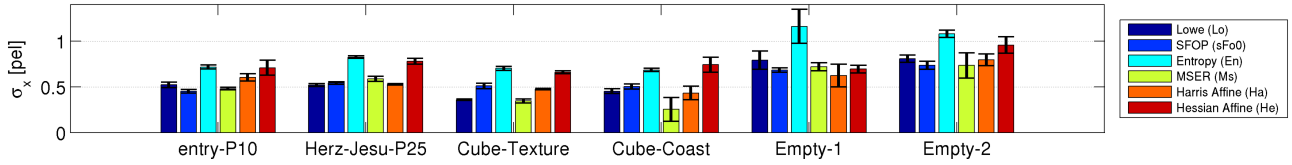

Fig. 5. Average estimated standard deviation $\widehat{\sigma}_{x'}$ of observations for individual detectors.

geometric model for the point location, as it is conceptually a window detector. For HESAF we believe that better accuracy could be achieved when using an improved subpixel localization method, as the points are conceptually similar to LOWE.

The accuracy of MSER features is noticeably strong on glCUBE-COAST and glCUBE-TEXTURE. This is especially interesting because a good performance of MSER on planar surfaces has been reported in other evaluations as well. It is also remarkable that the scores for sFOP are among the best ones on EMPTY-2 and EMPTY-1, although the other detectors did only calibrate part of the images here, usually the subset with less difficult texturings.

For combinations of detectors, the differences vanish due to the averaging.

Image Coverage. We see in Fig. 6 that sFOP and ENTROPY best cover the image with features in case of glCUBE-COAST, EMPTY-2 and EMPTY-1, which all show rather poor texturedness. On the other datasets they are slightly outperformed by MSER, while LOWE yields very similar results.

Accuracy of Estimated Camera Poses. Comparing the estimated projection centers to the ground truth poses for individual operators (Fig. 7 top), we see that the overall best results are achieved by sFOP and MSER, again with a special suitability of MSER for the planar surfaces. LOWE also yields very good results, but falls back on the smoothly textured glCUBE-COAST dataset, which also relates to the low number of object points achieved here (Fig. 3). Taking also into account its overall performance on EMPTY-2 and EMPTY-1, it seems that the LOWE detector is more suited for images with high texturedness. ENTROPY performs especially well on the HERZ-JESU-P25 dataset, which is quite surprising as neither the standard deviation $\widehat{\sigma}_{x'}$ nor the number of object points $\overline{N}_I$

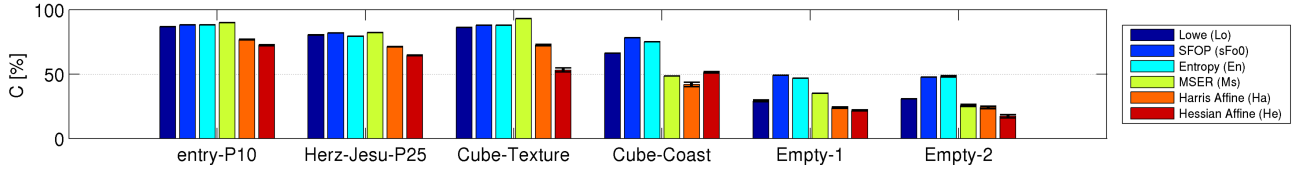

Fig. 6. Average area C [%] of the convex hull of image observations w. r. t. image area.

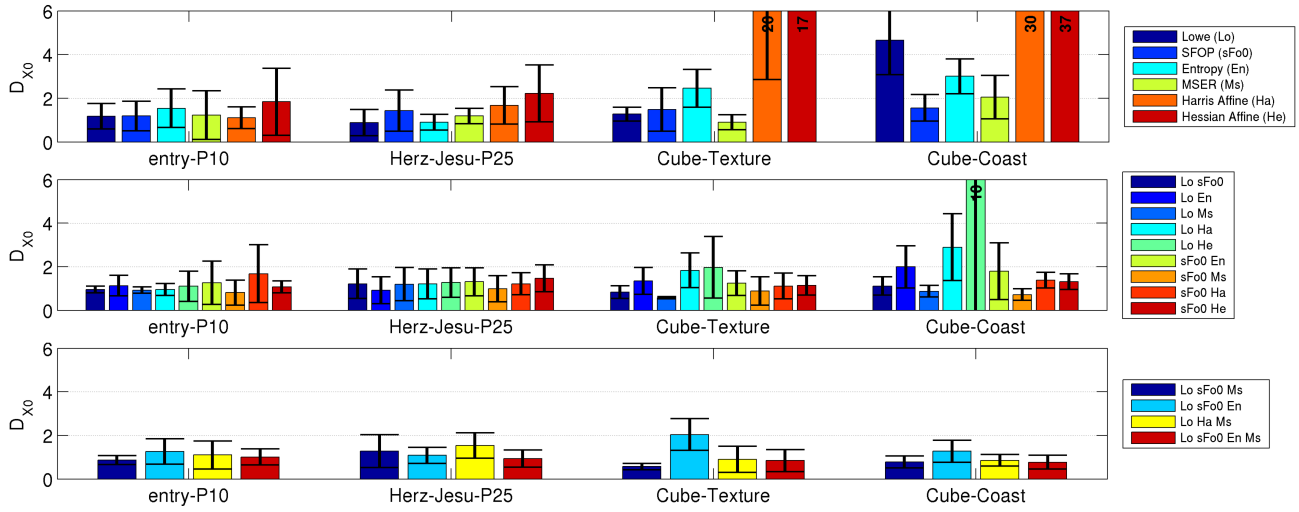

Fig. 7. Average squared distance $\overline{D}_{X_0}$ of reconstructed projection centers w. r. t. ground truth after a transformation into the coordinate system of the ground truth data.

was noticeable here. A reasonable explanation for this might be good geometric alignment of the features.

Combining detector pairs significantly improves the results, making them almost all acceptable. However, the pairwise combination of Lowe and Hesaf is again conspicuous: While the poor results for Hesaf and Haraf (see the top row of Fig. 7) are mostly compensated when combined with other detectors, combining Lowe and Hesaf does not seem to be beneficial, especially on glCube-Coast (see second row of Fig. 7 on the right). The triple combinations however are all very stable, but combining Lowe with sFop and Mser is noticeably the most promising setting.

5 Conclusion

The applied detectors showed quite different performance on the datasets. In particular, the EMPTY-2 and EMPTY-1 datasets with small amount of texture could not be successfully processed by most of the detectors individually, except by the proposed SFOP detector. The latter one showed overall best performance in the sense that it yielded good results on all indicators and datasets.

Under medium or high texturedness of the images, LOWE, MSER, SFOP and ENTROPY are all suitable operators for the orientation problem considered here. The MSER detector showed special strength on planar surfaces, where it delivered very good localization accuracy and repeatability. The HESAF and HARAF detectors however did not reach the same performance as other detectors in our setting. Especially HESAF gave rather weak scores under many indicators; however, from the close relationship to LOWE, we believe that an enhanced subpixel localization and non-maximum suppression might improve the results.

The ENTROPY detector showed worse localization accuracy on the datasets compared to others, but nonetheless yielded acceptable estimation results compared to the ground truth data which may be due to good geometric alignment of the points. This is also indicated by very good coverage of the image area with observations.

Using combinations of features solves most of the problems observed in the individual cases, especially allowing for complete successful orientations, with few exceptions. The overall best results are achieved when combining SFOP with LOWE, possibly complemented by the MSER or ENTROPY detector. Besides computational complexity, negative effects seem to occur only when combining very similar detectors like LOWE and HESAF, which are both based on the Laplacian. This suggests that one should account for the complementarity of feature detectors when combining them; a topic which has been recently addressed in [25].

References

1. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision **60** (2004) 91–110
2. Mikolajczyk, K., Schmid, C.: Scale and Affine Invariant Interest Point Detectors. International Journal of Computer Vision **60** (2004) 63–86
3. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. Image and Vision Computing **22** (2004) 761–767
4. Tuytelaars, T., Van Gool, L.: Matching Widely Separated Views Based on Affine Invariant Regions. International Journal of Computer Vision **59** (2004) 61–85
5. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: Exploring photo collections in 3d. In: SIGGRAPH Conference Proceedings, New York, NY, USA, ACM Press (2006) 835–846
6. Pollefeys, M., Koch, R., Vergauwen, M., Van Gool, L.: Automated Reconstruction of 3D Scenes from Sequences of Images. In: ISPRS Journal Of Photogrammetry And Remote Sensing. Volume 55(4). (2000) 251–267

7. Mayer, H.: Robust Least-Squares Adjustment Based Orientation and Auto-Calibration of Wide-Baseline Image Sequences. In: ISPRS Workshop BenCOS 2005, Bejing, China (2005) 11–17
8. Roth, D.G.: Automatic Correspondences for Photogrammetric Model Building. In: Proceedings of the XXth ISPRS Congress, Istanbul, Turkey (2004) 713–718
9. Läbe, T., Förstner, W.: Automatic Relative Orientation of Images. In: Proceedings of the 5th Turkish-German Joint Geodetic Days, Berlin (2006)
10. Strecha, C., von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U.: On Benchmarking Camera Calibration and Multi-View Stereo for High Resolution Imagery. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, Alaska (2008)
11. Martinec, D., Pajdla, T.: Robust Rotation and Translation Estimation in Multiview Reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07), Minneapolis, USA (2007)
12. Zach, C., Irschara, A., Bischof, H.: What Can Missing Correspondences Tell Us about 3D Structure and Motion? In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, Alaska (2008)
13. Harris, C., Stephens, M.J.: A Combined Corner and Edge Detector. In: Alvey Vision Conference. (1988) 147–152
14. Förstner, W., Gülch, E.: A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centres of Circular Features. In: ISPRS Conference on Fast Processing of Photogrammetric Data, Interlaken (1987) 281–305
15. Förstner, W., Dickscheid, T., Schindler, F.: Detecting Interpretable and Accurate Scale-Invariant Keypoints. In: 12th IEEE International Conference on Computer Vision (ICCV'09), Kyoto, Japan (2009)
16. Nister, D.: An Efficient Solution to the Five-Point Relative Pose Problem. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. Volume 26., Washington, DC, USA, IEEE Computer Society (2004) 756–777
17. Fischler, M.A., Bolles, R.C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Communications of the ACM **24** (1981) 381–395
18. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2004)
19. Lourakis, M.I.A., Argyros, A.A.: Design and Implementation of a Sparse Bundle Adjustment Software Library Based on the Levenberg-Marquardt Algorithm. Technical report, Heraklion, Crete, Greece (2004)
20. Förstner, W.: Local entropy of an image patch. Note (2009)
21. Moreels, P., Perona, P.: Evaluation of Features Detectors and Descriptors Based on 3D Objects. International Journal of Computer Vision **73** (2007) 263–284
22. Lazebnik, S., Schmid, C., Ponce, J.: A Sparse Texture Representation Using Local Affine Regions. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 1265–1278
23. Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06), Washington, DC, USA (2006) 2169–2178
24. Arun, K.S., Huang, T.S., Blostein, S.D.: Least-squares Fitting of Two 3D Point Sets. IEEE Transactions on Pattern Analysis and Machine Intelligence **9** (1987) 698–700
25. Förstner, W., Dickscheid, T., Schindler, F.: On the Completeness of Coding with Image Features. In: 20th British Machine Vision Conference, London, UK (2009)