RELIABLE IMAGE MATCHING WITH RECURSIVE TILING

D. Novák, E. Baltsavias, K. Schindler

Institute of Geodesy and Photogrammetry, ETH Zürich, 8093 Zürich, Switzerland – (david.novak, manos, konrad.schindler)@geod.baug.ethz.ch

KEY WORDS: Automation, Matching, Processing, Orientation, Reconstruction

ABSTRACT:

This paper presents a method to improve the robustness of automated measurements while also increasing the total amount of measured points and improving the point distribution. This is achieved by incorporating a tiling technique into existing automated interest point extraction and matching algorithms. The technique allows memory intensive interest point extractors like SIFT to use large images beyond 10 megapixels while also making it possible to approximately compensate for perspective differences and thus get matches in places where normal techniques usually do not get any, few or false ones. The experiments in this paper show an increased amount as well as a more homogeneous distribution of matches compared to standard procedures used nowadays.

1. INTRODUCTION

In the past decade several techniques have been introduced for automatic interest point detection and matching over wide baselines (i.e. large changes in viewing direction and/or scale). Blob detectors such as SIFT [Lowe, 2004] and SURF [Bay et al., 2008] or multi-scale corner detectors like Harris-Affine [Mikolajczyk et al., 2002] are used to extract feature points, and descriptors of their local neighbourhoods are attached to these points. The descriptors are then used to establish correspondences, usually with variants of (approximate) nearest neighbour matching [Lowe, 2004].

Detectors and descriptors are designed to be invariant to moderate changes in scale, rotation, and even affinity. This strategy has been very successful for low-resolution tasks like image retrieval [Csurka et al., 2004] and low-accuracy image orientation [Snavely et al., 2006]. However in high-accuracy applications such as photogrammetric surveying and industrial metrology it still has important limitations. In such applications one is interested in matching thousands of points in very high resolution images – typically 10-50 megapixels – and in convergent network geometry. However, under serious distortions (especially large scale difference) standard methods at best find a handful of correct matches. A further classic problem of wide-baseline methods is that repetitive textures are easily mismatched. This tendency is amplified by invariant descriptors, which suppress subtle differences between repetitive elements.

To the best of our knowledge few attempts have been made to remedy the problem of inaccurate automated orientation of images. Barazzetti et al. (2010) have shown that it is possible to improve the accuracy of current computer vision implementations by introducing robustness via photogrammetric methods, yielding satisfying results.

An obvious strategy to overcome the mentioned difficulties is to match iteratively, i.e. find a few correct correspondences with a highly robust estimator such as RANSAC, estimate the relative orientation, then match again under a loosely enforced epipolar constraint, update the orientation, and so forth. However this strategy is computationally costly because of the large amount of samples required to reliably and repeatedly estimate and refine the relative orientation in the presence of many outliers. In addition many incorrect matches are detected which are then discarded again. Furthermore, it is possible that the relative orientation does not yield a correct solution thus creating false epipolar lines. Additionally, this does not help much in terms of foreshortening because even if epipolar images are used, local distortions are still present.

Therefore, a hierarchical matching strategy is proposed in order to overcome limitations of wide-baseline matching and make it more usable for practical photogrammetry with high resolution images. This discussion is limited to the two-view case, and the two images are called the *source* and *target* image. Instead of directly estimating the relative orientation, the target image is recursively split into tiles. For each tile only a small number of very reliable matches are found with a restrictive nearest neighbour threshold. Then the tiles are warped with individually estimated projective transformation matrices (homographies) to better match the source image. After warping, each tile is split into four tiles again. As the tiles get smaller, the observed 3D structure on average deviates less and less from a planar patch, such that the homography becomes an ever better local approximation of the perspective distortion.

The tiling is more global than affine-invariant interest-point detection, since it operates in a top-down fashion and estimates the image distortion for larger neighbourhoods, while at the same time being more local than global geometric verification, which uses the matches from the entire image. As a side effect, the split into local tiles also circumvents the problem of matching the many thousands of points found in a high-resolution image in one step, which requires more memory than is typically available even on modern workstations.

The number of reliable matches increases with the decreasing extent of tiles because of the diminishing perspective distortion, and also because of the reduction of similar points from repetitive texture, which are then not found in the same tile any more. This yields a larger set of correspondences as well as a reduced amount of outliers. The algorithm has been implemented in a C++ program called “Gummireifen”. Experimental results on several datasets show a consistent improvement in matching performance compared to conventional wide-baseline matching, with an average increase of 2 – 3 times more matches. In addition a correctness of 95% or better can be achieved on average before any relative orientation is applied. With additional crude relative orientation the correctness of the matches rises above 98%, while the number of matches diminishes by less than 10%.

2. TILING

2.1 In a nutshell

The method can be explained intuitively by considering the observed 3D surface. If a digital surface model and the camera orientations were available, one could project the target image onto the surface and render it from the viewpoint of the source image, then the resulting synthetic image should perfectly match the source image pixel by pixel up to radiometric differences.

In the absence of the surface model, one can replace it with a projective transformation which is estimated from four point correspondences. Since the planar approximation of the surface induces distortions, it is integrated into a multi-level partition of local tiles, which each have their own projective transformation matrix. In this way, each tile of the target image at the lowest level is almost in correspondence with the source image and thus matching is more reliable.

The tiling and approximate local projective transformation per tile limit the search space for matching, which avoids many false matches. It improves matching robustness by reducing the perspective distortion of objects and reduces memory usage since each tile can be processed separately. Furthermore, it yields approximate camera positions for 3D point triangulation and iterative relative orientation from the projectivity estimated at the highest level.

2.2 Details

The images for the tiling approach are automatically preprocessed with a Wallis filter [Wallis, 1976]. Afterwards this method does not differ from standard wide-baseline matching: SIFT interest points are extracted independently in the two images, and matched with nearest-neighbour search and a nearest-neighbour threshold of 0.5. With the detected correspondences, a homography from the target to the source image is estimated with RANSAC [Fischler and Bolles, 1981]. Note that homography fitting is more efficient and also considerably more reliable than directly estimating the relative orientation. Not only does it require 4 correspondences instead of 5 (the minimal set exponentially influences the required number of samples), but more importantly the constraint has dimension 1 instead of 2 (i.e. the residuals are point-to-point distances, not point-to-line distances), such that incorrect fits are a lot less likely.

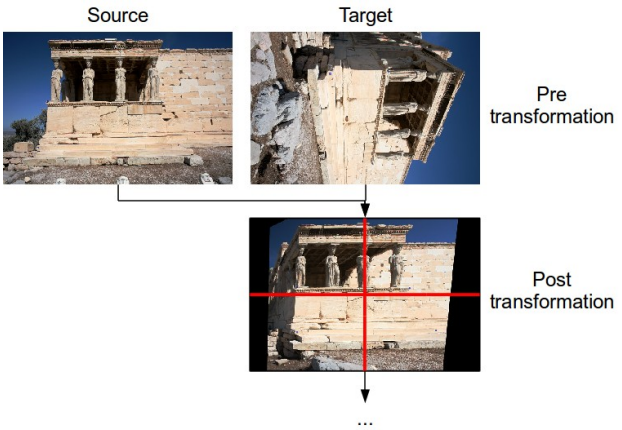


Figure 1: The first tiling step of the procedure.

With this homography obtained from the original images the target image is warped. Then the images are divided into four equal tiles. While for strongly non-planar surfaces the homography is only a rough approximation, it will nevertheless align the target image sufficiently well with the source image to ensure that the overwhelming majority of possible matches are between points in the same tile (e.g. the upper-left tiles of the source and target images contain approximately the same part of the 3D scene) as can be seen in Figure 1.

In the next step, one can thus repeat the procedure (interest point extraction / matching / homography estimation) for each tile independently, and then split each tile again into four sub tiles, thus generating a recursive grid of local projective alignments. The nearest-neighbour threshold is reduced to 0.4 to improve the overall correctness of the matches.

As the recursive tiling progresses, the projective alignment becomes a better approximation in most tiles, simply because the true object surface deviates less from a plane in a smaller neighbourhood (except on discontinuities). The natural stopping point would be when the 3D scene in one tile is perfectly planar, or tiles become so small that they no longer contain enough texture to find four good matches. This is obviously scene-dependent: scenes with large planes and/or large homogeneous areas will reach that point with bigger tiles. Empirically a final tile size of about 500 pixels has been determined to be sufficient for close-range photogrammetry applications, corresponding to 2 - 4 recursive splits. Note that at the lowest level it is not necessary to estimate homographies. For the last step a nearest neighbour threshold of 0.33 is chosen.

After projective alignment, the remaining distortions are so small that the matches can be verified with more discriminative, less invariant similarity measures to further reduce the number of matching errors. First, all correspondences are tested with normalized cross-correlation and those with a correlation coefficient smaller than 0.5 are discarded. Finally, the image coordinates of the remaining correspondences are refined with least-squares matching [Gruen, 1985] for further computations, and in the process those are discarded for which the matching does not converge, indicating that there is no reasonable affine transformation possible, which optimally aligns the patches. Before this step is finished, duplicate matches as well as ambiguous ones (i.e. two points share the same coordinate in the source image but have a different coordinate in the target image) are being eliminated.

After the image measurement step is complete, the measurements themselves are used to calculate a relative orientation between the two images to get rid of gross outliers. This is being done by exploiting the parameters given by the projective transformation. The projectivity of the first tiling step indicates the rough position of the target image relative to the source image. The base vector between the source and target image can thus be formulated as in (1), (2) and (3):

$$Bx = -x_0 - \frac{C_{11}\cdot x_0 + C_{12}\cdot y_0 + C_{13}}{C_{31}\cdot x_0 + C_{32}\cdot y_0 + C_{33}} \quad (1)$$

$$By = 1 - \left(\frac{A_{before}}{A_{after}}\right) \quad (2)$$

$$Bz = y_0 - \frac{C_{21}\cdot x_0 + C_{22}\cdot y_0 + C_{23}}{C_{31}\cdot x_0 + C_{32}\cdot y_0 + C_{33}} \quad (3)$$

C_{nn} denotes the elements of the projective transformation matrix, x_0, y_0 denote the centre of the target image. A_{before}, A_{after} are the area of the target image before and after the transformation.

Bx, By and Bz correspond to the [*X/Y/Z*] position of the target image relative to the source image which is given the [*X/Y/Z*] coordinates of [*0/0/0*]. The angular values ω, ϕ, κ for

the source and target image are initialized to $\frac{\pi}{2}/0/0$. With these starting values, the iterative relative orientation based on the coplanarity condition [Mikhail et al. 2001] is being performed. To avoid problems with outliers, five matches are randomly chosen and a relative orientation is calculated as part of a RANSAC procedure. All image measurements that have a residual that is larger than five pixel are excluded. It's possible that the approximations for the base vector are not correct (wrong sign), which causes the model to be flipped along the *X*-axis. This can be avoided by checking the *Y* values of the model coordinates and then correct the sign of the approximations accordingly.

3. EXPERIMENTS

3.1 Experimental set-up

In all experiments the same SIFT & nearest neighbour library [Hess, 2010] has been used (as included in the Gummireifen package). The exception is the Bundler [Snavely et al., 2006] test, where the nearest neighbour matching is being done by Bundler itself.

Each dataset has been processed several times: The 'standard' version is simple SIFT & nearest neighbour matching with a full resolution image. The nearest-neighbour threshold has been set to 0.33 and the image has been doubled in size [Lowe, 2004]. The 'standard-Wallis' version is the same as the 'standard' version except that it uses Wallis-filtered images as input. The 'tiling-meas' version only applies the tiling algorithm without the relative orientation and without the cross-correlation and least-squares matching refinement. The 'tiling-full' version applies all of the mentioned techniques above including a relative orientation for further outlier reduction. All tests have been performed on a 64-bit operating system thus allowing the normal SIFT procedure to use full resolution images. Each test shows the amount of matched points as well as the percentage of correct matches which have been investigated visually. Furthermore, a point distribution map shows the distribution of the points in one of the images for the 'standard' and the 'tiling-full' version. The point distribution map shows dark areas with less matched points and bright areas with more matched points.

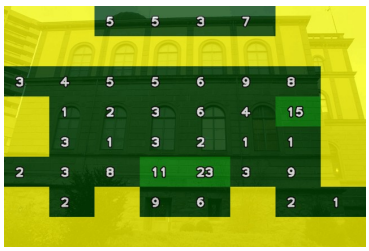
Furthermore a comparison has been done between Bundler, standard SIFT & nearest-neighbour matching and Gummireifen. All three versions generate image measurements which are imported into Bingo [Kruck, 1984] where relative orientation is computed with the RELAX [Kruck and Trippler, 1995] module and then bundle adjustment is performed in order to yield error statistics. The row with the number of points shows the amount of points going into the bundle adjustment and the amount of points that were accepted (out). For Bundler, the matched points before its internal bundle adjustment procedure have been used as number of input points. The amount of output points is obtained after the internal bundle adjustment and processing with Bingo.

If a dataset is marked as (calibrated) it means that additional camera parameters have been used in the Gummireifen package to correct the lens distortion. Since it is not possible to introduce additional parameters into Bundler manually they were not used – the same goes for the standard version.

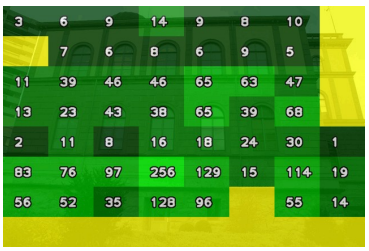
Sternwarte Zürich (calibrated)

	# matched points	% correct
standard	308	58.8
standard-Wallis	709	61.7
tiling-meas	2199	95
tiling-full	2049	99.8

standard



tiling-full

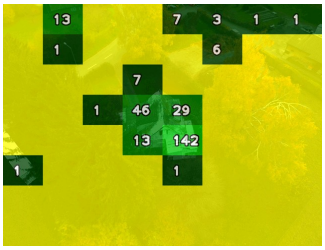
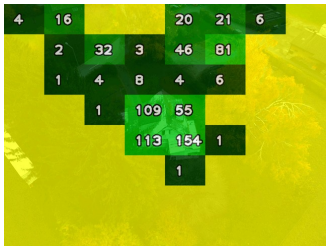


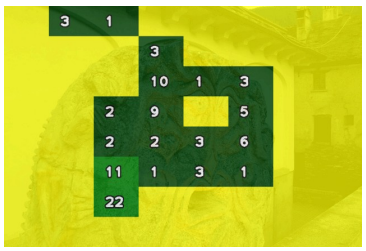
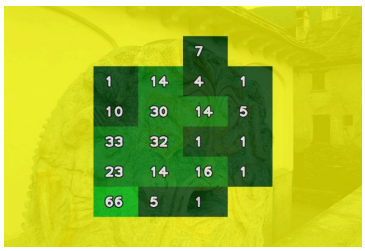
Bundle Adjustment Results

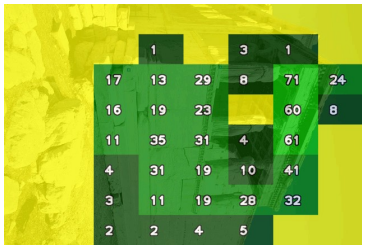
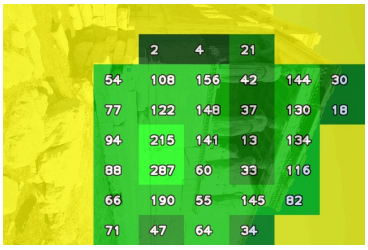
	# of points (in/out)	RMSE x/y (µm/px)
Standard	RO failed	RO failed
Bundler	209 / 118	2.93 / 0.53
Gummireifen	2049 / 1970	1.02 / 0.19

Octocopter over HXD (not calibrated)

	# matched points	% correct
standard	286	94.8
standard-Wallis	205	91.7
tiling-meas	708	98.4
tiling-full	688	100

standard	tiling-full	Bundle Adjustment Results		
			# of points (in/out)	RMSE x/y (µm /px)
		Standard	RO failed	RO failed
		Bundler	267 / 253	0.61 / 0.30
		Gummireifen	688 / 659	1.14 / 0.57

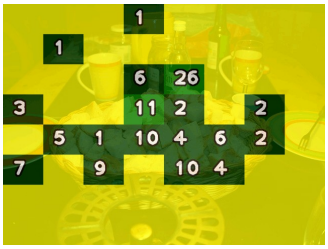
Bocca (calibrated)				
			# matched points	% correct
		standard	220	78.2
		standard-Wallis	88	87.5
		tiling-meas	284	98.6
		tiling-full	279	100
standard	**tiling-full**	**Bundle Adjustment Results**		
			# of points (in/out)	RMSE x/y (µm/px)
		Standard	220 / 169	4.64 / 0.76
		Bundler	192 / 170	3.01 / 0.49
		Gummireifen	279 / 278	1.00 / 0.16

Erechtheion (not calibrated)				
			# matched points	% correct
		standard	703	91.9
		standard-Wallis	662	92.1
		tiling-meas	3100	98.5
		tiling-full	3028	100
standard	**tiling-full**	**Bundle Adjustment Results**		
			# of points (in/out)	RMSE x/y (µm/px)
		Standard	703 / 635	8.98 / 1.10
		Bundler	669 / 593	5.38 / 0.66
		Gummireifen	3028 / 3006	9.26 / 1.13

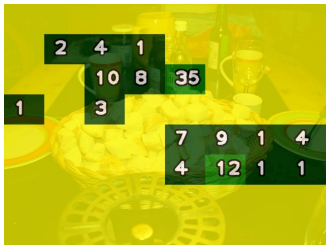
Pontresina bread basket (not calibrated)

	# matched points	% correct
standard	134	82.1
standard-Wallis	130	82.3
tiling-meas	123	83.7
tiling-full	107	96.3

standard



tiling-full



Bundle Adjustment Results

	# of points (in/out)	RMSE x/y (µm/px)
Standard	134 / 97	0.8 / 0.24
Bundler	110 / 90	1.20 / 0.35
Gummireifen	107 / 106	2.21 / 0.65

3.2 Relative orientation experiments

The relative orientation approximations, coming from the projectivity, are very helpful to ensure converging bundle adjustment. Experiments showed that more measurements do not necessarily mean that the possibility of multiple solutions in relative orientation is reduced. The approximations do help in this case to create a robust and correct relative orientation which can be later used in the bundle adjustment. A simple example of a two-image case is shown in Figure 2. Standard relative orientation without approximations fails in RELAX. When the approximations for the orientation parameters and model coordinates are introduced from the Gummireifen package, the relative orientation is successful.

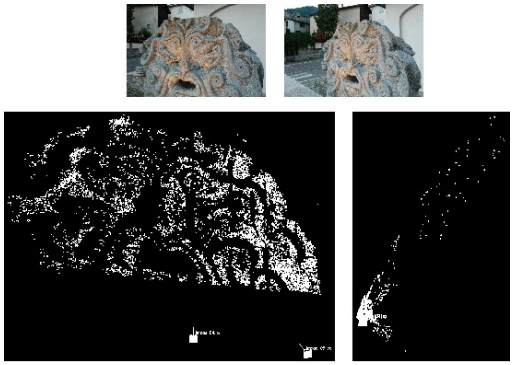


Figure 2:
Left: Successful relative orientation
Right: Failed relative orientation

4. DISCUSSION

The most obvious observation of the experiments is that except for the bread-basket dataset, the tiling approach achieves more matches than a standard procedure commonly used. Furthermore, the point distribution in the tiling approach is more homogeneous. If enough texture information is available in the images the tiling approach generates up to five times more matches in the overlapping area than a standard procedure. The standard procedure delivers generally a lower number of correct matches and lower correctness than Gummireifen, even with a restrictive nearest-neighbour threshold.

Two datasets are particularly interesting: the *Sternwarte* dataset and the *bread basket* dataset. In case of the *Sternwarte* the standard procedure delivers a rather poor performance which comes from the high amount of repeated texture and a convergence angle of roughly 30°. Due to this a lot of image points have been placed on similar looking object features which are matched incorrectly. The tiling keeps this under control.

The bread basket dataset is shown as a failure case and in which the projective transformation is estimated inaccurately. In this case, a lot of points have been found in two small areas. The transformation only fits well to these areas while the other areas are wrongly projectively transformed. This in turn means that the tiles no longer represent the same area of the object and that few matches with poor distribution are found. Furthermore, the dataset itself has the problem that the object-depth-to-distance ratio of the scene is large.

The usage of the Wallis filter seems to be generally beneficial as recommended in Baltsavias, (1991) but it is not fully consistent as can be seen in the *Octocopter* dataset where there is a higher amount of correct matches without the Wallis filter. This is most likely caused by the noise and the vegetation, found in the dark areas of this dataset. The Wallis filter enhances contrast especially in dark areas and the repetitive texture of trees and grass cause matching errors. Generally though it seems to be at least similar to the non-Wallis-filtered case with some cases clearly indicating an improvement.

The bundle adjustment results of Gummireifen might seem to contradict the claims of the paper because the *Octocopter*, *bread basket* and *Erechtheion* datasets deliver poorer performance than Bundler even though there is a very high correctness of the matches. This is caused by the fact that Bundler discards a lot of points near the border of the images, either because the feature extraction did not extract anything or because Bundler did discard them in its own bundle adjustment process. The accuracy in these areas is naturally lower due to an increased influence of the radial distortion an poorer ray intersection – all three datasets were used without calibration parameters. Figure 3 and Figure 4 illustrate this problem in the Erechtheion dataset. The confidence ellipsoids for the porch of the Caryatids in the foreground are roughly the same size in both procedures. However the error ellipsoids on the back wall are rather large and only show up in the Gummireifen case – since these points are situated near the border of the image, they

get influenced more by the radial distortion, leading to a higher RMSE value. The same happens in the Octocopter dataset where points far away from the house are obtained with Gummireifen and used in the adjustment while Bundler only keeps points near the centre of the image.

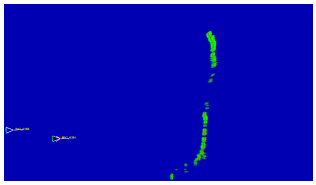


Figure 3: Bundler confidence ellipsoids

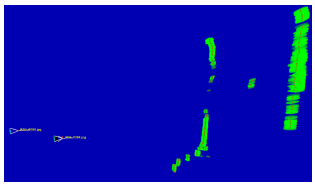


Figure 4: Gummireifen confidence ellipsoids

However, it is desirable to keep these border points to stabilize the image block when more images are used. If self-calibration is performed for Gummireifen on the Erechtheion dataset, the RMSE value drops to 1.27/0.16 µm/px. These values are however a bit optimistic since camera calibration with just two images is not recommended. One can see however in the two calibrated datasets that camera calibration can bring a significant improvement in terms of accuracy, as expected.

The fact that the standard case in the Erechtheion dataset delivers slightly better results than Gummireifen stems from the fact that fewer points have been found on the back wall, thus reducing the influence of the radial distortion and the RMSE. The standard version fails to perform relative orientation in two cases. For the Sternwarte it is easily explainable since there is no pre-processing done thus the outlier rate of over 40% causes RELAX to fail. In the *Octocopter* case the small baseline causes a flip of the camera positions and the relative orientation fails as well.

The main drawback of the tiling technique is speed. Usually it takes about three times longer to measure points than the standard case. If the cross-correlation and least-squares matching as well as the relative orientation are included, the processing time takes up to five times longer than that of the standard procedure.

5. CONCLUSION & OUTLOOK

The experiments show that the tiling approach creates more matches with higher correctness than a standard procedure. However, a price has to be paid for the increased density and correctness which means that time-critical applications can not benefit from it very much.

Further improvements can be made to increase robustness. One of them would be to adjust the tile size/shape adaptively depending on the point distribution and image content. Furthermore, a non-linear warping function instead of the projective transformation might be used to avoid the failure cases.

As has been shown in Barazzetti et al. (2010) it is possible to improve the image measurements to produce results that do not have an error of loop closure as seen in Bundler. Error of loop closure is a common problem in structure from motion software when walking around a building or recording the walls of a plaza. Further investigations will be done to see, if the technique proposed in this paper can achieve similar results. Once the software can handle multiple images it will be interesting compare the accuracy to existing packages. It is expected that the overall more homogeneous point distribution and the approximations for the relative orientation will allow for a robust and accurate orientation.

REFERENCES

Baltsavias, E. P., 1991, *Multiphoto geometrically constrained matching*, Ph.D. Dissertation, Institute of Geodesy and Photogrammetry, ETH Zürich, Mitteilungen No. 49.

Barazzetti, L., Scaioni, M., Remondino, F., 2010, Orientation And 3D Modelling From Markerless Terrestrial Images: Combining Accuracy With Automation, *The Photogrammetric Record*, 24(132):356 – 381.

Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., 2008, SURF: Speeded Up Robust Features, *Computer Vision and Image Understanding*, 110(3): 346 – 359.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., Bray, C., 2004, Visual categorization with bags of keypoints, *Proceedings in Workshop on statistical learning in computer vision*, ECCV, pp. 1 – 22.

Fischler, M. A., Bolles, C. R., 1981, Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography, *Comm. of the ACM,* 24: 381 – 395.

Gruen, A. W., 1985, Adaptive Least Squares Correlation: A Powerful Image Matching Technique, *South African Journal of Photogrammetry, Remote Sensing and Cartography*, 14: 175 – 187.

Hess, R., 2010, An Open Source SIFT Library, *In Proceedings ACM Multimedia(MM).*

Kruck, E., 1984, BINGO: A Program for Bundle Adjustment for Engineering Applications – Possibilities, Facilities and Practical Results, *International Archives of Photogrammetry and Remote Sensing,* Comm. V, XXV (A5): 471 – 480.

Kruck, E., Trippler, S., 1995, Automatic Computation of Initial Approximations for Bundle Block Adjustment with Program RELAX*, Optical 3-D Measurement Techniques III,* Vienna, October 2 – 4.

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

Mikhail, E. M., Bethel, J. S., McGlone, J. C., 2001, *Introduction to Modern Photogrammetry*, John Wiley & Sons, New York.

Mikolajczyk, K., Schmid, C., 2002, An affine invariant interest point detector, in *Proceedings of the 7th European Conference on Computer Vision*, Copenhagen, Denmark.

Schaffalitzky, F., Zisserman, A., 2003, Automated Location matching in movies, *Computer Vision and Image Understanding*, 92(2): 236 – 264

Snavely, N., Seitz, S., M., Szeliski, R., 2006, Photo Tourism: Exploring image collections in 3D, *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2006).*

Wallis, R., 1976, An approach to the space variant restoration and enhancement of images, *Proceedings of Symposium on Current Mathematical Problems in Image Science*, pp. 329 – 340.