# ISSUES FOR IMAGE MATCHING IN STRUCTURE FROM MOTION

**Helmut Mayer**

Institute of Geoinformation and Computer Vision, Bundeswehr University Munich
Helmut.Mayer@unibw.de, www.unibw.de/ipk

**KEY WORDS:** Computer Vision, Matching, Accuracy Analysis, Machine Vision, Close Range Photogrammetry, Sensor orientation

## ABSTRACT:

This paper discusses issues for matching point features in structure from motion, i.e., pose estimation and 3D reconstruction. As our building block are image triplets for which matches can be checked also geometrically, we discuss the issues for triplets based on partial or final results of a complete system for structure from motion. Particularly, we analyze and try to give answers to the following three questions: Shall all good matches be used or only the best matches? Does covariance information help? and finally – Is least-squares matching useful? We discuss the issues also from a theoretical point of view, but we mostly address them empirically by means of three suitable, i.e., difficult examples, as we found that answers to the above questions are strongly tied to the characteristics of real data.

## 1 INTRODUCTION

Structure from motion or pose estimation and 3D reconstruction or also short orientation relies on the correspondence of features in the images. In most cases point features are matched, because points fix two degrees of freedom.

This paper discusses several issues for matching points for structure from motion, especially important for wide-baseline setups. They are particularly concerned with answering the following three questions:

- Shall all good, i.e., also ambiguous, matches be used or only the best matches?
- Does covariance information help?
- Is (relative) least-squares matching useful or are (sub-pixel) point coordinates sufficient?

Our structure from motion procedure (cf. also (Mayer, 2005)) starts by generating image pyramids, extracting Förstner (Förstner and Gülch, 1987) points at a resolution of about 100 × 100 pixels and matching them by means of cross correlation. To deal with rotations in the image plane, we use the eigenvector of the structure tensor employed in the Förstner operator to normalize the orientation of the patches before cross correlation.

All matches with a cross correlation coefficient (CCC) above an empirically found loose threshold of 0.5 are then checked a second time based on affine least-squares matching. For the resulting matches a much more conservative threshold of 0.8 is used for CCC and additionally only matches with an estimated accuracy below 0.1 pixel are retained. Though we are able to deal with uncalibrated data based on fundamental matrix and trifocal tensor and a successive direct self-calibration, i.e., no approximate values are needed, following (Pollefeys et al., 2004), we have resorted to (at least approximately) calibrated cameras lately. All examples of this paper rely on calibrated cameras and for them we employ the 5-point algorithm (Nistér, 2004). To deal with wrong matches, we use Random Sample Consensus – RANSAC (Fischler and Bolles, 1981), the Geometric Robust Information Criterion – GRIC (Torr, 1997), and after computing 3D structure robust bundle adjustment (McGlone et al., 2004).

Once epipolar lines are known for image pairs, they are used to restrict the search space when orienting triplets on the next highest level of the image pyramid. As there is no good and accepted solution for the direct orientation of calibrated triplets yet, we employ two times the five point algorithm, once from image one to two and once from image one to three, both times fixing the coordinate system and the rotations for the first camera. I.e., the only information still unknown is the relative scale of the two 3D models, which we determine as median of the distance ratios to 3D points generated from homologous image points. For images beyond one Megapixel the triplets are oriented a second time on the third highest level of the pyramid, this time taking into account the information about the orientation of the triplet in the form of the trifocal tensor for matching.

For the triplets of a sequence orientation can be computed independently from each other. This fact can be used to speed up processing by distributing the different pairs and triplets to the available processor cores. First experiments with a quad-core have shown a speed up of more than three compared to using only a single core.

Once all triplets have been oriented, they are linked based on the overlap between them and new points in appended triplets are added to the set of points. To speed up the linking, it is done hierarchically, leading from triplets to quadruples, sextuples, tentuples, eighteen-tuples, etc. (double number minus an overlap of two). The points of the oriented sequence are finally tracked to the original image resolution. Examples show that this works also for tens of images.

Concerning former work (Mikolajczyk et al., 2005) and (Mikolajczyk and Schmid, 2005) are of special interest. In (Mikolajczyk et al., 2005) different affine region detectors are compared concerning their performance when changing viewpoint, scale, illumination, defocus and image compression. (Mikolajczyk and Schmid, 2005) compares the next step, namely local descriptors, such as SIFT (Lowe, 2004) or PCA-SIFT (Ke and Sukthankar, 2004) concerning affine transformation, scale change, image rotation, image blur, JPEG, and illumination changes. While all these issues are important, they do only partially address the issues posed in this paper. Particularly, (Mikolajczyk et al., 2005) and (Mikolajczyk and Schmid, 2005) are both not interested into the possibility to select a correct solution based on the achieved accuracy, which we found to be a very efficient means for solving challenging structure from motion problems. This is particularly addressed by our second and third question where we present the effects of using highly accurately matched points and even covariance information. While (Mikolajczyk and Schmid, 2005) shows, that in general SIFT-based descriptors work best, they found that the CCC works nearly as well for a large overlap of the patches, which we enforce by least-squares matching.

The remainder of the paper is organized according to the given three issues. Thus in Section 2 we discuss if all good or only the best matches shall be used. Section 3 addresses the importance of covariance information and Section 4 the advantage of using sub-pixel point coordinates versus relative coordinates generated via least-squares matching. The paper ends up with conclusions. Statistics given for the examples below pertains to ten trials each if not stated otherwise.

## 2 ALL GOOD VERSUS BEST MATCHES

When matching point features between images the problem arises, that there can be more than one valid match in terms of a matching criterion such as thresholds on the CCC or the estimated accuracy of the match of a given point in one image to a point in another image. A typical example where this problem arises are images of facades with one type of window where, e.g., all upper left corners, can be matched very well to another in the other image. As one is moving when capturing the images, particularly for windows behind the facade plane for which parts are occluded depending on the perspective, the best match might even be exactly the wrong window after moving.

In (Lowe, 2004) and (Mikolajczyk and Schmid, 2005) dealing more with object recognition it is recommended to use a match if it “unambiguous”, i.e., if there is no similarly good match measured by a ratio. We have found in numerous tests that for our typical data with many facades of buildings this strategy is inferior to a strategy where one uses only one, namely the best, match for each point, which additionally has to be evaluated beyond a given threshold, as there are often too few unambiguous matches even for very high thresholds for the ratio.

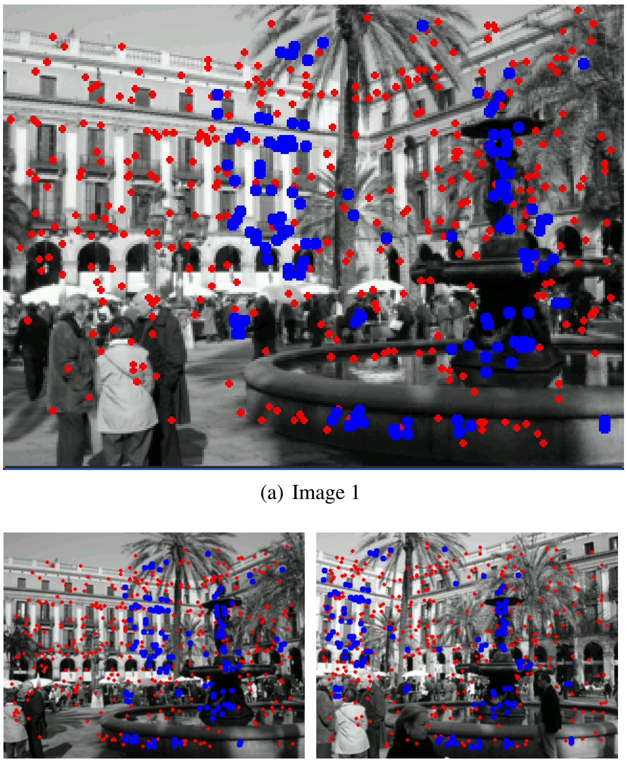
In this paper we analyze a specific variety of this problem: Often best matches are only checked in one direction using one image as ‘master” and the others as “slave” images, with multiple matches in the slave images not being detected. We show the scale of this problem for image triplets and also its possibly severe effect on 3D reconstruction.

Tab. 1 gives statistics for multiple matches for the three example triplets used throughout this paper. The results are after least-squares matching with thresholds for the CCC of 0.8 and for the estimated accuracy of 0.1 pixel, with the best matches taken for images 1 to images 2 and 3. While for the triplets Desk (cf. Fig. 5) and House (cf. Fig. 6) the number of multiple matches from the second or third image into the first image is relatively low, for triplet Real (cf. Fig. 1) showing the Plaza Real in Barcelona Spain, 187 out of 761 or nearly 25% of the points in images 2 and 3 map to two points in image 1. Additionally, there is a lower, but still not neglectable number of points which map to three and even up to five points in image one. Additionally, 52 of the matches are multiple matches in image 2 as well as 3.

Fig. 1 shows the correct result for triplet Real if only the best matches for all images, i.e., no multiple matches at all, are employed. The analysis of the effect of using only the best or all good matches is done for the triplet Real for the calibrated triplets on the second highest level of the pyramid as they are central for solving the structure from motion problem: Once matches have been checked by intersecting three rays for each point, in most cases the orientation is either correct and the refinement is straightforward, or the orientation is plainly wrong and there is nothing to be done about it.

| Triplet | Real | Desk | House |
|---|---|---|---|
| # matched points | 761±7 | 331±6 | 238±10 |
| double matches | 187±8 | 377±3 | 39±2 |
| triple matches | 36±3 | 4±1 | 6±1 |
| quadruple matches | 5±1 | 0 | 1±1 |
| quintuple matches | 2±1 | 0 | 0 |
| Double multi-matches | 52±3 | 5±1 | 4±1 |

Table 1: Statistics for many-fold matches for three image triplets



(a) Image 1

(b) Image 2

(c) Image 3

Figure 1: Triplet Real taken with 5 Megapixel Sony P100 camera – Correct solution for “best matches only”. Accepted matches are given as big blue points while all other points are shown as red smaller points.

To gain more insight into the problem, we have devised three experiments additionally to our standard strategy of using the best matches, i.e., no multiple matches at all, for all images, namely using **all good matches**, i.e., all matches with CCC $>$ 0.8 and estimated accuracy $<$ 0.1 pixel after least-squares adjustment:

- for triplets, i.e., calibrated trifocal tensors, on the second highest pyramid level only,
- for point pairs, i.e., essential matrices, on the highest pyramid level only, and finally
- for pairs as well as triplets.

In all cases we match on the second highest level of the pyramid 1896 points in image 1, where we employ a local maximum suppression scheme, against 4972 points in image 2 and 5604 points in image 3. Tab. 2 shows the results. Basically, for our standard strategy of best matches nearly every time the correct result as checked by visual inspection of the resulting 3D model is obtained. On the other hand, when using all, i.e., multiple, matches in images 2 and 3 for pairs 1 and 2 and 1 and 3 as well as for the triplet, we never obtained any correct result at all. The accepted matches in Tab. 2 are the best matches produced by RANSAC for the triplet with the best RANSAC solutions for the test run in terms of their GRIC value polished by robust bundle adjustment.

| matches | given matches | accepted matches | $\sigma_0$ | failure rate |
|---|---|---|---|---|
| best | 471±4 | 128±16 | 0.12±0.02 | <1% |
| all triplets | 755±9 | 136±25 | 0.14±0.03 | 5% |
| all pairs | 439±5 | 88±13 | 0.11±0.02 | 10% |
| all | 777±8 | 151±13 | 0.16±0.01 | 100% |

Table 2: Statistics for triplet Real on the second highest level of the pyramid. Our standard strategy of using only the best matches is compared to employing all possible matches for the triplet on the second highest pyramid level, for the pairs on the highest pyramid level, as well as for both. $\sigma_0$ is the average standard deviation after robust bundle adjustment.

Things are more complex if one just uses the best matches only for pairs or only for triplets. The statistics are a little bit worse for pairs than for triplets. Here, the image triplet Real is of particular interest because it shows the rare case of a repetitive structure conspiring with the orientations of the cameras in a way, that two very similar solutions can be obtained, only one of which is correct. More precisely, the clearly wrong solution in Fig. 2, where points at the windows on a large part of the left facade in images 1 and 2 are mapped to points on a small part of the same facade in image 3 has a very similar evaluation in terms of the GRIC value as the correct solution. While the correct solution consists of 70 matches with an average standard deviation after robust bundle adjustment $\sigma_0$ of 0.141 pixel, the incorrect solution is made up of 67 matches with a $\sigma_0$ of 0.144 pixel.

When using the best matches in the triplets only, the situation is a little bit better. Yet, solutions as shown in Fig. 3 are neither optimal nor stable as the 3D structure is mostly determined by the fountain in the foreground.

Note: Very often random simulations help to find errors in the modeling or in the program. Yet, sometimes errors occur due to regularities in the data which are unlikely to be reproduced by simulation. In the above case these are the regularities and self-similarities of windows on facades which do not only lead to wrong matches, as all windows look very similar, but also to wrong configurations matching close-by windows in two images to more distant windows in the third image.

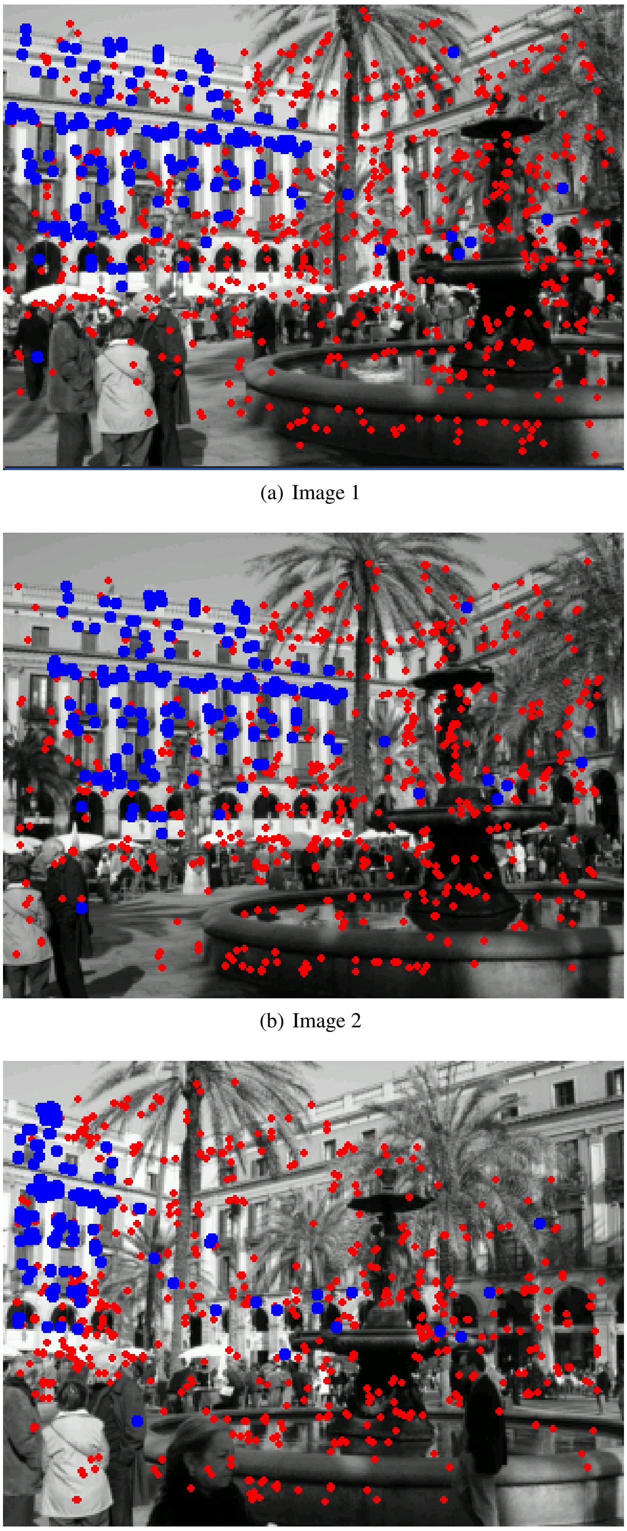
## 3 VARIANCES VERSUS FULL COVARIANCE MATRICES

The second question we address pertains to the usefulness of full covariance matrices

$$\begin{bmatrix} \sigma_x & \sigma_{xy} \\ \sigma_{yx} & \sigma_y \end{bmatrix}$$

describing error ellipses (cf. Fig. 4) versus reduced covariance matrices

$$\begin{bmatrix} (\sigma_x + \sigma_y) \cdot 0.5 & 0 \\ 0 & (\sigma_x + \sigma_y) \cdot 0.5 \end{bmatrix} .$$


(a) Image 1

(b) Image 2

(c) Image 3

Figure 2: Incorrect solution for "all good matches for pairs" with a GRIC value very close to a correct solution – points cf. Fig. 1.

We employ the covariance information in two ways: First, we use them in the bundle adjustments for pairs and triplets. Additionally, we improve with them the selectivity of the test in RANSAC whether a point is an inlier. We basically use Phil Torr's Geometric Robust Information Criterion – GRIC (Torr, 1997) in the form of the ratio of the squared residual $e$ between measured and reprojected point and the average squared error of the point $\sigma$,

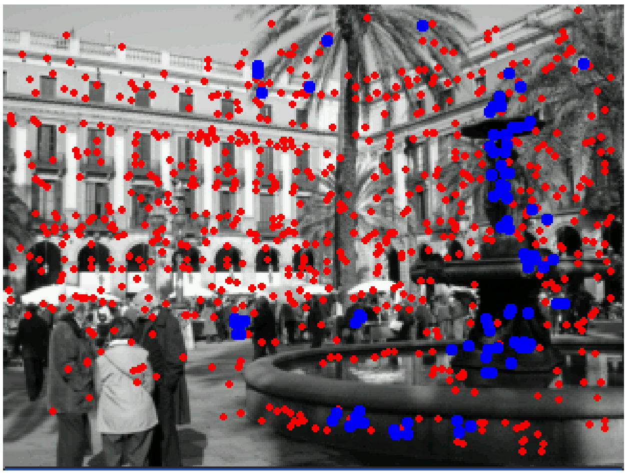


Figure 3: Correct, but sparse and unstable solution, as most matches are on the fountain in the foreground, for “all good matches for triplet” – points cf. Fig. 1

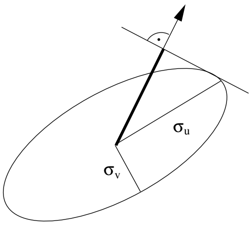


Figure 4: Error ellipse modeled by full covariance matrix with large and small semi-axes $\sigma_u = \sqrt{\lambda_1}$ and $\sigma_v = \sqrt{\lambda_2}$, $\lambda_1$ and $\lambda_2$ being the eigenvalues of the covariance matrix. For a given difference vector the part of the error in its direction as determined by the intersection of its normal tangential to the ellipse is shown in bold.

i.e., $\frac{e^2}{\sigma^2}$. But instead of the average error, we compute from the particular error ellipse the part of the error in the direction of the difference vector (cf. Fig. 4).

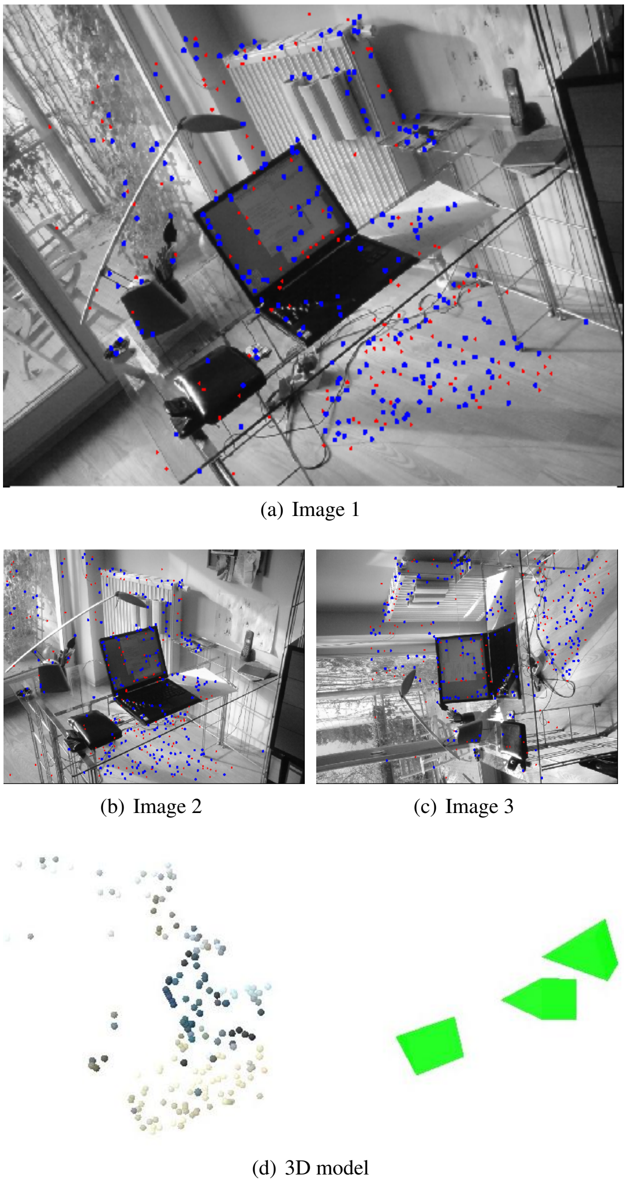
Obviously, full covariance information for the 2D points is only useful if the error ellipse is not circular. Also, if the ellipse is aligned with the coordinate axes and the direction of movement, which is the case for forward or sideward moving triplets or sequences, its influence on the result is still low. This is especially true as we only extract points with a not too elliptical error ellipse, i.e., points which are not too line-like.

In Tab. 3 statistics are given for the ratio of the semi-axes of the error ellipses $\sigma_u$ and $\sigma_v$ (cf. Fig. 4) for all three example triplets of this paper. The minima show, that there is basically no absolutely circular ellipse. Yet, the rest of the values leads to the conclusion that most ellipses are not extremely elliptical with averages below or even well below 2 for partly very large maxima (we note that the maxima are for hundreds of points).

| $\sigma_u/\sigma_v$ | $\mu$ | $\sigma$ | min | max |
|---|---|---|---|---|
| House | 1.69 | 0.77 | 1.05 | 41.2 |
| Desk | 1.56 | 0.65 | 1.05 | 66.1 |
| Real | 1.97 | 0.93 | 1.03 | 97.9 |

Table 3: Statistics for the ratio of the semi-axes $\sigma_u$ and $\sigma_v$ of the error ellipse (cf. Fig. 4)

To show the possible impact of the covariance information, we use a scene taken with a low-quality digital camera in a mobile phone (Sony Ericson K550) with a 2 Megapixel Sensor and a lense diameter of about 2 mm. We rotated the camera two times with about 45°. Fig. 5 shows the triplet as well as a visualization of the 3D result. Even though the quality of the images is relatively low due to the weak contrast of the camera with its tiny objective, the result is still very reliable.



(a) Image 1

(b) Image 2

(c) Image 3

(d) 3D model

Figure 5: Triplet Desk taken with rotated 2 Megapixel Sony Ericson K550 mobile phone camera – points cf. Fig. 1; green pyramids in the 3D model symbolize the camera positions.

Tab. 4 shows a comparison in the form of several characteristic values for the case of reduced and full covariance matrices, this time for the triplets on the third highest level of the pyramid, i.e., points before tracking, and after tracking the points down to the original 2 Megapixel resolution. One can see that when using the full covariance matrices not only significantly more points are accepted and tracked down to the original resolution, but also the average standard deviation $\sigma_0$ of the final bundle adjustment is slightly better.

## 4 POINT COORDINATES VERSUS LEAST-SQUARES MATCHING

The final question we address concerns the usefulness of the additional information from least-squares matching compared to sub-

| | reduced | full covariance |
|---|---|---|
| # points before tracking | 192±10 | 227±3 |
| # points after tracking | 178±8 | 206±3 |
| $\sigma_0$ [pixel] | 0.45±0.01 | 0.43±0.01 |

Table 4: Statistics for triplet Desk employing full and reduced covariance information

pixel coordinates from point extraction according to (Förstner and Gülch, 1987).

More specifically, we analyze the influence of the improvement of the relative coordinates by means of least-squares matching of 9 × 9 patches in the reference image employing affine, i.e., six parameter, transformation for the geometry and bilinear interpolation for resampling. To do so, we compare on the second highest level of the pyramid the orientation of triplets for a really wide-baseline triplet (cf. Fig. 6(e) view from top). The basic statistics can be found in Tab. 5. It shows, that the least-squares matching results into more points with a much better accuracy.

| | matched points | $\sigma_0$ |
|---|---|---|
| Förstner points only | 28±3 | 0.27±0.03 |
| + least-squares matching | 41±8 | 0.11±0.01 |

Table 5: Statistics for triplet House employing only the coordinates of the Förstner points or additionally also least-squares matching

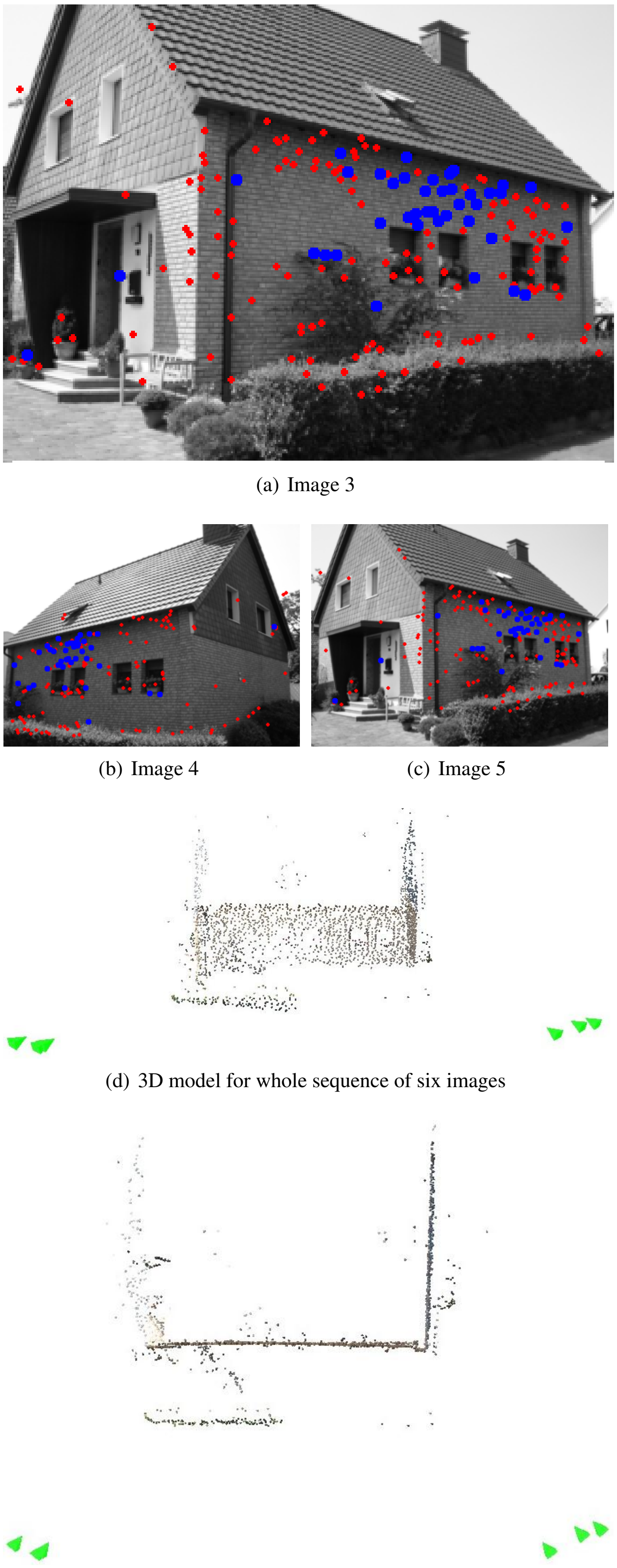
The high accuracy of the matched points also translates into a meaningful solution (cf. Fig. 6). On the other hand, Fig. 7 shows a typical result if only the Förstner points are employed. It is deteriorated in a way that it is not useful any more. In not much more than 10% of the trials an at least approximately correct result could be obtained, comprising still only 40 to 50 points compared to on average about 125 points for the correct solution for the triplet. The average standard deviation of the final bundle adjustment is 0.5 to 0.7 pixel, compared to 0.16 pixel for the correct solution using relative coordinates produced by least-squares matching.

Concerning the correct result please note, that one deficit of our approach is, that it is not scale invariant. This means, the matching only works for those regions in the images, where the scale is similar, i.e., the scale difference is smaller than 20 or 30%. Yet, this deficit stems only from the basic correlation step. The least-squares matching could (naturally) also deal with larger variations, as long as it obtains an estimate of the scale difference. At least we have shown that we can cope with in-plane rotations (cf. example Desk above).

The 3D model for the whole sequence (cf. Fig. 6(d) and 6(e)) consisting of six images shows, that even for this complex wide-baseline setup it is possible to obtain fully automatically a highly accurate 3D representation. It consists of 2186 3-fold, 115 4-fold, 134 5-fold, and 14 6-fold points and the average standard deviation $\sigma_0$ is 0.15 pixel.

## 5 CONCLUSIONS

We firstly have shown that it can be important to use only the best matches and not all possible matches. This is mostly due to the ambiguity of multiple matches.



(a) Image 3

(b) Image 4

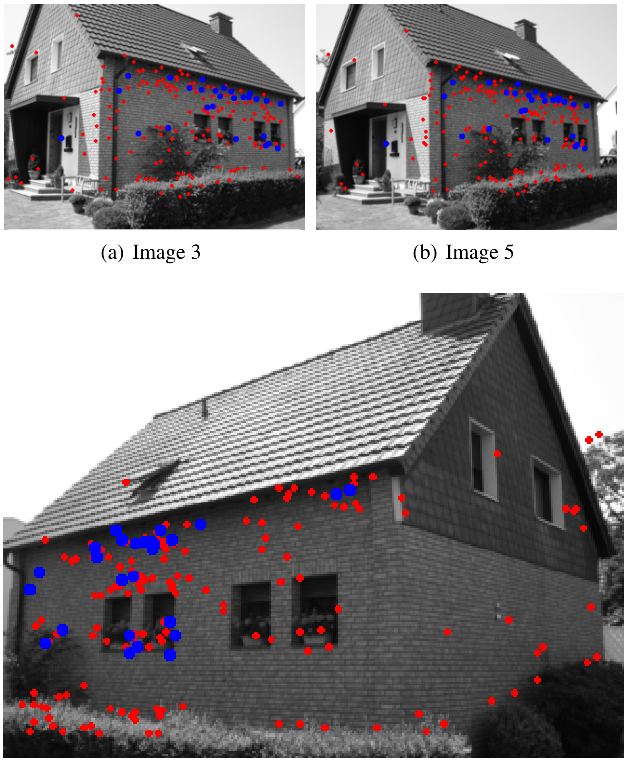
(c) Image 5

(d) 3D model for whole sequence of six images

(e) 3D model – view from top

Figure 6: Sequence House taken with 5 Megapixel Sony P100 camera – Correct result for images 3, 4, and 5 when employing least-squares matching – for points and pyramids cf. Fig. 5

While we have presented evidence, that the 2D covariance information from least-squares matching can be helpful for matching, we note, that this is valid only when the images are rotated relative to each other. When employing least-squares matching it is not much effort to consistently use the covariance information throughout the process for the acceptance of matches as well as for bundle adjustment, but also the gain is often not very big.



(a) Image 3

(b) Image 5

(c) Image 4

Figure 7: Incorrect result for images 3, 4, and 5 of Sequence House without least-squares matching; the error is mainly in image 5 – for points cf. Fig. 1

On the other hand, this does not pertain to 3D covariance information, which can also be obtained reliably from reduced 2D covariance matrices. 3D covariance can be extremely useful particularly for smaller baselines where the accuracy in the depth direction can be much worse than in the other two directions. For decisions such as the determination of planes from 3D points based on RANSAC, it can be very helpful to employ 3D covariance information.

We have finally shown, that for more difficult examples it is not enough to use the coordinates of points, even if they are sub-pixel precise. Thus, we recommend to use least-squares matching as the final step of the determination of homologous points.

Concerning future work we particularly think about dealing also with scale differences. Here, the SIFT feature extractor (Lowe, 2004) is very attractive due to its speed. Once an estimate of scale (difference) and orientation is available, we plan to use our least-squares matching approach. Additionally, our success with determining the rotation for the Förstner point operator has inspired us to think about using it also to determine the relative scale. While scale-normalization can be done by means of scale-space theory (Lindeberg, 1994), the issue is efficiency.

## REFERENCES

Fischler, M. and Bolles, R., 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Communications of the ACM 24(6), pp. 381–395.

Förstner, W. and Gülch, E., 1987. A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centres of Circular Features. In: ISPRS Intercommission Conference on Fast Processing of Photogrammetric Data, Interlaken, Switzerland, pp. 281–305.

Ke, Y. and Sukthankar, R., 2004. PCA-SIFT: A More Distinctive Rrepresentation for Local Image Descriptions. In: Computer Vision and Pattern Recognition, Vol. 2, pp. 516–523.

Lindeberg, T., 1994. Scale-Space Theory in Computer Vision. Kluwer Academic Publishers, Boston, USA.

Lowe, D., 2004. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), pp. 91–110.

Mayer, H., 2005. Robust Least-Squares Adjustment Based Orientation and Auto-Calibration of Wide-Baseline Image Sequences. In: ISPRS Workshop in conjunction with ICCV 2005 "Towards Benchmarking Automated Calibration, Orientation and Surface Reconstruction from Images" (BenCos), Beijing, China, pp. 1–6.

McGlone, J., Bethel, J. and Mikhail, E. (eds), 2004. Manual of Photogrammetry. American Society of Photogrammetry and Remote Sensing, Bethesda, USA.

Mikolajczyk, K. and Schmid, C., 2005. A Performance Evaluation of Local Descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(10), pp. 1615–1630.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T. and van Gool, L., 2005. A Comparison of Affine Region Detectors. International Journal of Computer Vision 65(1/2), pp. 43–72.

Nistér, D., 2004. An Efficient Solution to the Five-Point Relative Pose Problem. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(6), pp. 756–770.

Pollefeys, M., Van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K. and Tops, J., 2004. Visual Modeling with a Hand-Held Camera. International Journal of Computer Vision 59(3), pp. 207–232.

Torr, P., 1997. An Assessment of Information Criteria for Motion Model Selection. In: Computer Vision and Pattern Recognition, pp. 47–53.