Performance evaluation of the five-point relative pose with emphasis on planar scenes[1)]

Siniša Šegvić, Gerald Schweighofer and Axel Pinz
EMT, TU Graz
axel.pinz@tugraz.at, http://www.emt.tugraz.at/~vmg/

Abstract:
We consider performance evaluation of the state-of the-art solution for recovering the relative pose between two calibrated views. Our focus is on planar scenes which are not tractable by algorithms which do not enforce the so-called calibrated constraint. The capability to cope with planar scenes has therefore been stressed as an important advantage of the novel five-point algorithm. However, we show that for planar and nearly planar scenes there is a considerable degradation of the five-point algorithm performance under noise. This is especially the case for sidewise motion, for which substantially better motion hypotheses can be obtained by homography decomposition. The differences are even greater when more than five points are available, since the accuracy of the homography approach scales better. We also note that, contrary to the previous claims, the five point algorithm is not a method of choice even in non-planar overconstrained contexts, since the performance of the classical 8pt algorithm can be greatly improved by equilibration. Thus, the results imply that the five-point algorithm is the best option only for non-planar scenes in minimal cases (as a hypothesis generator in a RANSAC scheme). At the price of a perhaps acceptable performance deterioration, the five point algorithm could be used for planar scenes as well, but only for prevalently forward motion.

1 Introduction

We consider the problem of recovering the relative pose [5] (or relative orientation [6]) between the two images of a nearly planar scene. The relative pose consists of rotation and translation (up to an unknown scale factor) relating the metric coordinates of the two camera frames. It is known that two views of a plane can be explained by up to two relative poses, and that the ambiguity arises when all imaged points are closer to one of the two cameras [7]. This ambiguity can not be resolved without further information such as a third view of the scene or a priori knowledge about the orientation of the imaged plane.

[1)] This work has been supported by the European MC IIF project AViCMaL, the French national project Predit Mobivip, and the Austrian science fund FWF S9103. The authors would like to thank Arnold Irschara and Albert Diosi for fruitful discussions and helpful suggestions.

In general, the relative pose can be recovered only if the images have been acquired with calibrated cameras, allowing the points to be expressed in normalized coordinates. In this calibrated context, the epipolar constraint gives rise to the essential matrix, which yields a unique decomposition into the sought motion parameters [8]. Previous research predominantly addressed the more general projective or uncalibrated context, resulting in algorithms which could recover some geometric information even in images acquired by arbitrary cameras. The most widely known among these is the eight-point linear algorithm [6], which still provides competitive performance, especially when the procedure is properly equilibrated. Equilibration of a linear system can be viewed as a superset of normalization [4], and corresponds to multiplying the matrix of the system with appropriately chosen weight matrices from both sides [10]. However, the uncalibrated algorithms can not be applied to planar scenes, even when the cameras are calibrated [4]. This is because, for planar scenes, the epipolar constraint is satisfied by an infinite number of matrices, regardless of the number of correspondences [8]. The occurrence of this multiplicity of solutions is often referred to as planar degeneracy [15, 1, 3]. The dimension of the solution space can be lowered in the calibrated context by complementing the epipolar constraint with the enforcement of the characteristic algebraic structure of the essential matrix. This involves solving of a system of cubic equations which has been achieved only recently [11]. The main advantage of the new algorithm is that it can be applied to subsets of only five correspondences. In a typical random sampling environment, this ensures faster guessing in presence of outliers and, together with a fair execution speed, can significantly release the computational burden in a real-time application.

The five-point algorithm can return up to 10 motion hypotheses. A disambiguation can be performed by looking at the reprojection error of a sixth point [12]. In the ambiguous planar case without noise, two of the above hypotheses will perfectly satisfy all available constraints. However, this ambiguity would need to be addressed in any approach, so that the five-point algorithm has been recommended even in the case of planar scenes [11, 14]. The proposed disambiguation approach relies on a third view of the scene [11]. This solution is attractive due to its generality: nearly the same algorithm can be used both in cases of planar and volumetric scenes. However, *the accuracy of the recovered relative pose in the presence of ambiguity* has not been appropriately investigated in the previous research. This paper aims at filling that void, by comparing the five-point algorithm to the approach based on the homography decomposition. Additionally, an independent evaluation of other experiments from [11] is performed and the discovered discrepancies are brought to attention.

2 The planar ambiguity

It is widely known that there is a linear relation between the corresponding homogeneous [9] points q_{i1} and q_{i2} in two images of a planar scene [7]. The resulting transformation $\mathbf{P}^2 \rightarrow \mathbf{P}^2$ is known as homography, and can be represented as a 3×3 matrix $\mathbf{H}$, such that $\mathbf{H} \cdot q_{i1} = q_{i2} \, \forall i$. The homography is uniquely defined by the geometry of the two cameras ($\mathbf{R}$,$\mathbf{T}$), and the normal $\mathbf{n}$ and the distance d defining the plane $\mathbf{n}^\top \mathbf{x} = d$ in the frame of the first camera:

$$\mathbf{H} = \mathbf{R} + \frac{1}{d}\mathbf{T} \cdot \mathbf{n}^\top \tag{1}$$

The reverse procedure is also possible. Using the recovered homography between the two sets of corresponding points, one can inquire about the relative pose and the geometry of the imaged plane. Unfortunately, the solution is not unique since each homography generates 8 decomposition hypotheses [2]. By enforcing the visibility constraint for all point correspondences (i.e. that the target is in front of both cameras), one arrives to at most two physically valid hypotheses. The ambiguity occurs only if all of the observed points are closer to one of the two cameras [7]. This occurs quite often in practice, making the proper treatment of the ambiguous case compulsory. The ambiguity is always present for predominantly forward motion (which is characteristic for navigation), but it may even occur for sidewise motion depending on the position of the planar target.

The planar ambiguity is illustrated on an experiment involving a noiseless artificial planar dataset. Figure 1 shows two images of 9 co-planar points giving rise to two concurrent reconstructions. Assume that the physically correct solution is the one at the left side of the figure, and let the front camera move in a circle around the back camera in a way that the target remains visible. Then the ambiguity ceases to exist when the front camera reaches a position in which some points of the target become closer to the back camera. The concurrent decomposition then becomes noticeably wrong, since the reconstruction of the target "rips apart" in a way that the "discriminating" points are reconstructed behind the both cameras.

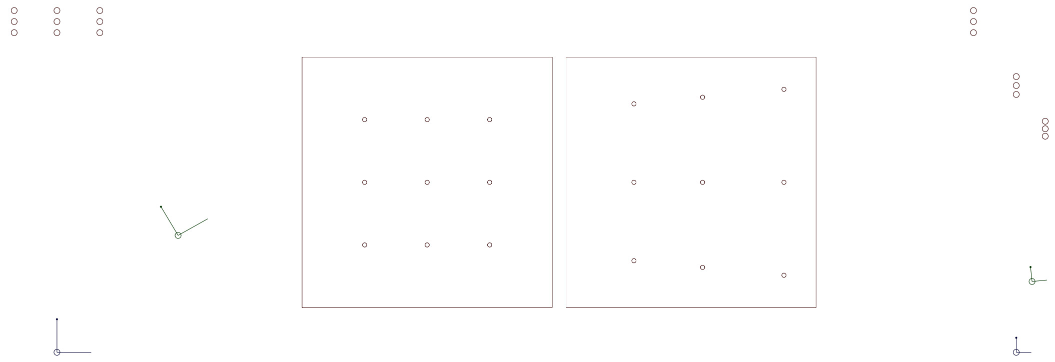


Figure 1: Two images of 9 planar points (middle) and the two reconstructions (left, right).

3 Experimental results

The recovery of the relative pose is usually performed in the following three steps: (i) rejection of erroneous correspondences using random sampling, (ii) re-estimation on the set of inliers, and (iii) iterative refinement by bundle adjustment. Note that the second step is also very important, since the success of the bundle adjustment greatly depends on a correct initial solution [13]. The analysis is therefore performed for (i) minimal cases with exact solutions, and (ii) overconstrained cases where the noise in input data is reduced by redundancy.

3.1 The experimental setup

The experiments have been performed on artificial noisy data. The employed setup is similar to [11], but additional specifications are provided in order to ensure the repeatability of the results. For each camera, the referential left-oriented coordinate system is set so that the image plane is defined by the equation $z = 1$. For both cameras the horizontal field of view is 45°. The geometry of the camera pair is defined by the angle θ, which defines the translation direction in the x-z plane. The rotation of the 2nd camera around the common y axis is set in a way that its optical axis passes through the centroid of the imaged points. Thus, $\theta = 0$ implies forward motion with no rotation. The random point cloud is instantiated in a volume visible by both cameras, placed between two parallel planes perpendicular to the optical axis of the first camera. The distance from the first camera to the closer plane of the volume is always 10 baselines, while the distance between the planes (`depth`) is varied between 0 and 5 baselines. The distribution of the points in the volume is uniform. The points are projected to the two images, and both image coordinates are perturbed with a zero-mean Gaussian noise [4]. The standard deviation of the noise is expressed in pixels of a 352×288 image as $\sigma = 1.0$.

The experiments involve 10000 applications to random samples of point correspondences. The multiple hypotheses provided by the 5pt algorithm are first disambiguated by requiring that all of the reconstructed 3D points be in front of both cameras. In the minimal cases, the 5pt algorithm is tested on samples of six points, where the sixth point is used to select the best among the surviving hypotheses [12]. The selection can be based either on the reprojection or on the Sampson error [4], with similar suboptimal but acceptable results. In the overconstrained cases, near optimal disambiguation can be achieved by looking at overall reprojection error induced by a triangulation scheme assuming the error in one image [11].

The experiments address the distribution of the angular error in the recovered translation direction as the harder part in recovering the relative pose [10, 11]. As in [11], experiments with minimal cases consider the first quartile of the error distribution $q_1\{\Delta T\}$, while the median value $med\{\Delta T\}$ is used to characterize overconstrained cases. The experiments were performed in Matlab and C++ using the implementations of the five-point algorithm provided by the authors[1], and within the library VW34[2] from the University of Oxford, respectively.

[1] `http://vis.uky.edu/~stewe/FIVEPOINT/`

[2] `http://www.doc.ic.ac.uk/~ajd/Scene/Release/vw34.tar.gz`

3.2 Comparison of the 5pt and 8pt algorithms for non-planar scenes

The results obtained for non-planar scenes with `depth`=5 are presented in Figure 2, depending on the groundtruth translation direction. The algorithm labeled `5pt-ideal` shows the results obtained by taking into account the *best* hypothesis from each sample of five correspondences: this provides a notion of the success of the disambiguation schemes relying on the 6th point (minimal cases) and the overall reprojection error (overconstrained cases). In the minimal cases, the results confirm the previous findings that the 5pt algorithm is a better option than the 8pt algorithm. However, the results reported in [11] (figure 12, middle) approximately match only for forward motion: our results for sidewise motion are 100% worse, while the extremum at 50° is 50% higher. In our experimental setup, a notable improvement for $\theta \in (30°, 90°)$ has been noticed for an enlarged field of view and the same pixel size.

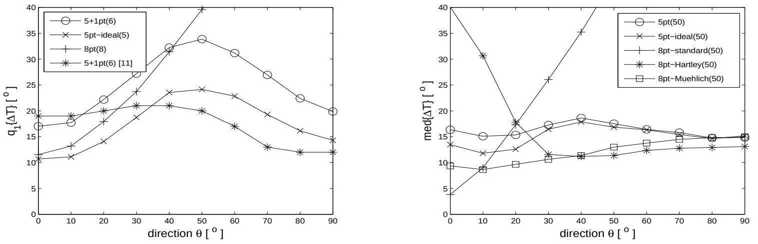


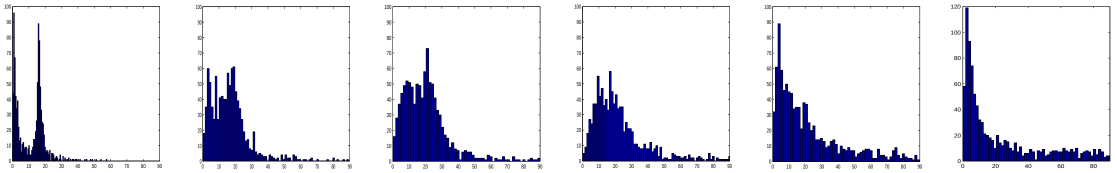
Figure 2: Translation error ΔT in degrees plotted against the groundtruth translation direction θ for `depth`=5, in minimal ($q_1\{\Delta T\}$, left) and overconstrained cases (50 points, $med\{\Delta T\}$, right).

In the overconstrained cases, the 5pt algorithm is compared with standard (`8pt-standard`) [6], normalized (`8pt-hartley`) [4] and equilibrated (`8pt-muehlich`) [10] 8pt algorithm. The results disprove the claims from [14], that the 5pt algorithm is the most consistent option when more points are available, since the equilibrated 8pt is better or equal for all translation directions.

3.3 Five-point algorithm and planar scenes

In the presence of planar ambiguity, the five-point algorithm produces both feasible motions, among other hypotheses. The presented disambiguation schemes often show little preference among the two motions, resulting in a loss of about 50% hypotheses, as illustrated in Figure 3 for minimal cases. The figure shows that the resulting distributions are bimodal only for unrealistically small noise. When the scene gains depth, the mode of the distribution corresponding to the correct motion (ΔT=0) becomes more and more distinct.

To assess the adequacy of the 5pt algorithm for planar scenes, in Figure 4 we compare it to the specialized solution based on the decomposition of the planar homography (`hg`). Both contexts are addressed, with respect to whether the additional information required for resolving the planar ambiguity is available or not. The former "ideal" context is simulated in algorithms



depth=0, σ=0.02 depth=0, σ=0.1 depth=0, σ=0.2 depth=1, σ=0.2 depth=5, σ=0.2 depth=10, σ=0.2

Figure 3: Frequency distribution of the translation error (ΔT) for the 5pt algorithm with 6 points and θ=15°. Number of observations (ordinate) is plotted against ΔT (abscissa) in the interval $[0, 90°]$. The varied parameters are depth of the scene and noise in pixels σ.

`5pt-ideal` and `hg-ideal`, by taking into account the best among the returned hypotheses. In minimal cases, it makes sense to compare algorithms operating on the same number of points: `5+1pt(6)` and `hg(6)`, and `5pt-ideal(5)` and `hg-ideal(5)`. The results suggest that the homography approach is hindered by the problems similar to those presented in Figure 3. Nevertheless, the homography generally provides superior performance except for the forward motion where it is only slightly better.

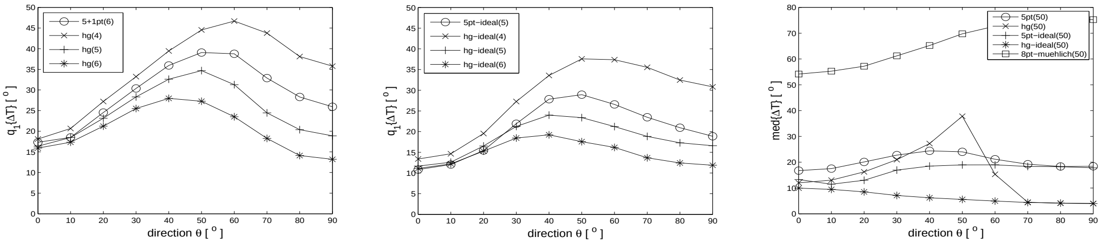


Figure 4: Results for planar scenes in minimal cases (left), minimal cases with ideal disambiguation (middle), and overconstrained cases (right).

It is straightforward to verify that, in the considered experimental setup, the planar ambiguity can not be resolved for directions $|\theta|$ less than approximately 64.9°. Then a corner of the artificial volume becomes equidistant from the two cameras, and the points which are closer to the first camera can be used for disambiguation [7]. This direction also roughly corresponds to the greatest difference between the "regular" and ideal algorithms, implying the greatest extent of the planar ambiguity. In the overconstrained cases, one *must* resort to additional information in order to ensure useful results for the ambiguous configurations which occur so often. This can be seen in Figure 4 (right), where only chance and numerical predisposition define the exact position of the median of a "regular" overconstrained algorithm between the two modes of the distribution. Nevertheless, the conclusion is much clearer than in the minimal cases. The experiments suggest that the homography takes much better advantage of the additional points, making its advantage indisputable. An optimal general procedure therefore would need to rely on a criterion for optimal model selection such as reprojection error or [15]. Finally, Figure 5, illustrates that the above figures do not change significantly

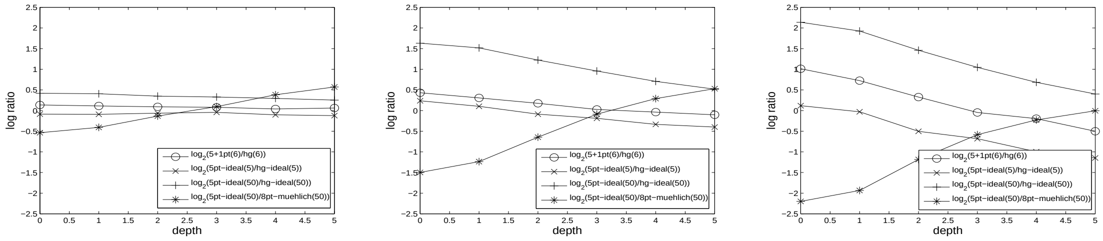


Figure 5: Log-ratio of the accuracy against the depth. Left: θ=0°, middle: θ=45°, right: θ=90°. As before, we consider $q_1\{\Delta T\}$ and $med\{\Delta T\}$ in minimal and overconstrained cases, respectively.

even for near-planar scenes. In the case of homography, the deviations are equivalent to additional noise and result in graceful performance degradation. The two approaches level-off approximately at `depth`=2 (minimal cases) and `depth`=4 (overconstrained cases).

3.4 Results for other parameters of the experimental setup

The presented experimental setup has been chosen in order to be able to compare our results with the previous work [11]. Many other experiments have been performed and qualitatively similar results have been obtained, while some variation trends have been identified. Experiments with other noise parameters $\sigma = \{0.2, 0.5, 1.5\}$ have shown that the relative performance of the homography with respect to the 5pt algorithm increases with noise, while the opposite is true for the equilibrated 8pt algorithm. The performance of the homography approach for forward motion increases for different orientations $(30°, 60°)$ of the target plane. The performance of the homography for sidewise motion decreases for a severely slanted target $(60°)$. Smaller target distance (2,5 baselines) results in improvement of the 8pt algorithm and deterioration of the homography approach. In all overconstrained experiments, the 5pt algorithm was outperformed either by homography or the equilibrated 8pt algorithm. Homography was better than the 5pt algorithm in all experiments with minimal cases and planar scenes.

4 Conclusion

The paper addressed performance evaluation of the recent five-point algorithm for recovering the relative pose. The results differ from what has been claimed in previous research, and imply that the five point algorithm is definitely not a preferable solution in the overconstrained context. In the minimal context, the five-point algorithm is a method of choice for non-planar scenes and can also be successfully applied to planar scenes, though a homography with the same number of points is likely to score better, especially for sidewise motion. The homography approach tends to provide a more accurate result in the planar case, since it generates a unique solution, which accounts for both feasible motions. The epipolar geometry on the other hand gives rise to two essential matrices, which amplifies uncertainty in the presence of noise. Additionally, the homography fully constrains the mutual position between the two point matches, while the epipolar geometry is limited to the distance from the epipolar line. The above is confirmed by the experiments which show that the differences in accuracy are strongly correlated with the extent of planar ambiguity (the mutual distance between the two solutions), but persist even in configurations in which the duality is absent. The obtained experimental results suggest that the choice of the algorithm for recovering the relative pose is highly context dependent. It therefore seems that a general solution should make an attempt to choose the best among the three options (`8pt`, `5pt`, `hg`). The design of the appropriate tests in the calibrated context is an open area for the future research.

References

[1] Ondrej Chum, Tomás Werner, and Jiri Matas. Two-view geometry estimation unaffected by a dominant plane. In *Proc. of CVPR*, pages 772–779, San Diego, CA, USA, 2005.

[2] Olivier D. Faugeras and Francis Lustman. Motion and structure from motion in a piecewise planar environment. *Int. J. of Pattern Recog. and Artificial Intell.*, 2(3):485–508, 1988.

[3] Jan-Michael Frahm and Marc Pollefeys. Ransac for (quasi-)degenerate data (qdegsac). In *Proc. of CVPR*, pages 453–460, New York, NY, USA, 2006.

[4] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, Cambridge, UK, 2004.

[5] Berthold K. P. Horn. Relative orientation. *Int. J. Comput. Vis.*, 4(1):59–78, 1990.

[6] Hugh Christopher Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.

[7] Hugh Christopher Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Proc. R. Soc. London*, B227(1249):399–410, 1986.

[8] Y. Ma, S. Soatto, J. Košecká, and S.S. Sastry. *An Invitation to 3-D Vision: From Images to Geometric Models.* Springer-Verlag, New York, USA, 2004.

[9] Roger Mohr and Bill Triggs. Projective geometry for image analysis. A tutorial given at the International Symposium of Photogrammetry and Remote Sensing, Vienna, July 1996.

[10] Matthias Mühlich and Rudolf Mester. The role of total least squares in motion analysis. In *Proc. of ECCV*, pages 305–321, Freiburg, Germany, 1998.

[11] David Nistér. An efficient solution to the five-point relative pose problem. *IEEE Trans. PAMI*, 26(6):756–770, 2004.

[12] David Nistér. Preemptive ransac for live structure and motion estimation. *Mach. Vision. Appl.*, 16(5):321–329, 2005.

[13] Gerald Schweighofer and Axel Pinz. Fast and globally convergent structure and motion estimation for general camera models. In *Proc. of BMVC*, pages 147–157, Edinburgh, Great Britain, September 2006.

[14] H. Stewénius, C. Engels, and D. Nistér. Recent developments on direct relative orientation. *ISPRS. J. Photogramm.*, 60(4):284–294, June 2006.

[15] Philip Torr. An assessment of information criteria for motion model selection. In *Proc. of CVPR*, pages 47–53, San Juan, Puerto Rico, 1997.