Down to Earth: Using Semantics for Robust Hypothesis Selection for the Five-Point Algorithm

Andreas Kuhn[1,2], True Price[2], Jan-Michael Frahm[2], Helmut Mayer[1]

[1] Bundeswehr University Munich
[2] The University of North Carolina at Chapel Hill

Abstract. The computation of the essential matrix using the five-point algorithm is a staple task usually considered as being solved. However, we show that the algorithm frequently selects erroneous solutions in the presence of noise and outliers. These errors arise when the supporting point correspondences supplied to the algorithm do not adequately cover all essential planes in the scene, leading to ambiguous essential matrix solutions. This is not merely a theoretical problem: such scene conditions often occur in 3D reconstruction of real-world data when fronto-parallel point correspondences, such as points on building facades, are captured but correspondences on obliquely observed planes, such as the ground plane, are missed. To solve this problem, we propose to leverage semantic labelings of image features to guide hypothesis selection in the five-point algorithm. More specifically, we propose a two-stage RANSAC procedure in which, in the first step, only features classified as ground points are processed. These inlier ground features are subsequently used to score two-view geometry hypotheses generated by the five-point algorithm using samples of non-ground points. Results for scenes with prominent ground regions demonstrate the ability of our approach to recover epipolar geometries that describe the entire scene, rather than only well-sampled scene planes.

1 Introduction

Large-scale Structure from Motion (SfM) has tremendously progressed in the last decade [24, 7, 1, 18, 13, 23]. The key approach in SfM-type methods, and especially in incremental SfM, involves building up scene geometry from initial two-view relationships. To accomplish this, virtually all systems estimate epipolar geometry to establish the overlap between image pairs. When Exchangeable Image File Format (EXIF) information is available for the images, such geometry is typically obtained via essential matrix estimation using the five-point algorithm [21]; alternatively, the eight-point algorithm [11] can be used to estimate a fundamental matrix for image pairs with unknown calibration. To provide robustness against incorrect point correspondences, this estimation is usually placed in a RANSAC framework, where a minimal number of correspondences are iteratively sampled to provide statistical guarantees on the computed transformation. The estimated epipolar geometries are vital for recovering relative pose between the two images, which enables the triangulation of scene geometry and other subsequent steps in the SfM pipeline.

In this work, we address a key problem of RANSAC-based essential matrix estimation that arises when the available point correspondences in two images are heavily biased toward a single plane in the scene. We show that while the existing methods work reliably for noise-free data, essential matrix estimation often achieves only a local optimum on real-world data given noisy point detection and, importantly, an uneven 3D spatial distribution of the available point correspondences. We find that such errors often manifest themselves in 3D urban reconstructions, with computed essential matrices largely over-fitting to fronto-parallel structures, such as building facades, and mis-characterizing undersampled planes, in particular the ground. Indeed, for the vast majority of scenes, the popular large-scale reconstruction systems (*e.g.*, [28, 23]) almost always fail to capture the ground structure, especially for reconstructions from uncontrolled photo-collections. To account for this apparent limitation in essential matrix estimation, we propose a new approach for scoring proposals for two-view geometries, taking into account semantic labeling of the detected image features. Focusing on ground/non-ground labelings, we demonstrate that our method is able to correctly characterize the epipolar geometry in a variety of image pairs that were only partly correctly treated in a traditional RANSAC framework.

2 Background and Related Work

Solving for relative geometry from five point correspondences does not *per se* provide a globally optimal solution [10]. In general, there exist ten discrete solutions [5], which can be reduced only in special configurations by means of the five-point algorithm [21] itself. Even though global optimally algorithms have been presented [12] for solving for the essential matrix, they are unproven and no practical evidence for an efficient implementation is given [25].

It has long been recognized [26, 21, 3] that the validity of estimated epipolar geometry (calibrated or uncalibrated) is inherently tied to both the 3D spatial distribution and 2D accuracy of the 2D correspondences shared between a given image pair. As Nistér notes in his seminal five-point algorithm paper [21], however, calibrated two-view geometry estimation (*i.e.*, essential matrix estimation) enjoys substantially less ambiguity than its uncalibrated counterpart. Given perfect correspondences, the geometry provided by the five-point algorithm is unique, except in the case of a planar point set, where possibly a single ambiguity could arise. This is much more manageable than the case for fundamental matrix estimation, wherein a planar point set causes degeneracy [3, 31, 15]. Given its significant advantages and superiority in practice [22], calibrated two-view geometry estimation has largely been thought to be a solved problem.

The specific issue we aim to address is the case where essential matrix estimation in a RANSAC framework fails to find a correct solution due to a sub-optimal

3D spatial sampling of feature points in the image, which in the limiting case lie on a single scene plane. In our experience, noise and outliers in the sampled point sets can drive RANSAC to select geometry that overfits to the dominant scene plane(s) if not enough points on less-sampled planes, namely the ground, are available. We briefly explain our observations in the following, before introducing our proposed method.

3 Error Behavior for Oversampled Planes in the RANSAC-based Five-Point Algorithm

In this section, we provide a high-level overview of the error behavior associated with epipolar geometry estimation in a RANSAC framework for image pairs having planar-biased point sets. Our goal is to emphasize the importance of taking into account all essential scene planes when choosing the best solution from the hypotheses generated by this algorithm.

3.1 The Five-point Algorithm

Consider a pair of calibrated cameras observing a scene, and let K and $\bar{K}$ represent their intrinsic matrices. The essential matrix E describes the epipolar geometry of the cameras: $\bar{p}^T \bar{K}^{-T} E K^{-1} p = 0$, where p and $\bar{p}$ are corresponding 2D points in each image, respectively. Geometrically speaking, $\bar{K}^{-T} E K^{-1} p$ maps point p in the first image into an epipolar line in the second image, and the equality to zero constrains $\bar{p}$ to lie along this line. Relative camera pose (3-DOF rotation and 2-DOF translation) can be recovered, in part, by decomposing the essential matrix, which makes its estimation vital to tasks such as SfM [9].

Using Nistér's five-point algorithm [21], we can solve for E using just five point correspondences. The algorithm is formed from ten cubic constraints that are well-known properties of the essential matrix [9]:

$$\det(E) = 0,\ 2EE^T E - \operatorname{tr}(EE^T)E = 0\ . \tag{1}$$

When the constraints in Eq. (1) are used in conjunction with five point correspondences, the problem of solving for E can be reduced to finding the roots of a tenth-degree polynomial [21, 16]. This results in up to 10 possible essential matrices, all of which constitute a valid epipolar geometry for the original five correspondences. To select the correct solution for E, traditional methods leverage additional point correspondences and choose the solution with the largest support. Accordingly, the selection of the prevailing hypothesis from the up-to-ten possible essential matrices hinges on the remaining available correspondences.

3.2 Error Behavior in Epipolar Geometry Estimation

In real-world applications, image correspondences are typically obtained using feature matching, which often results in a large number of putative correspondences that are affected by noise and outliers. To robustly recover relative camera

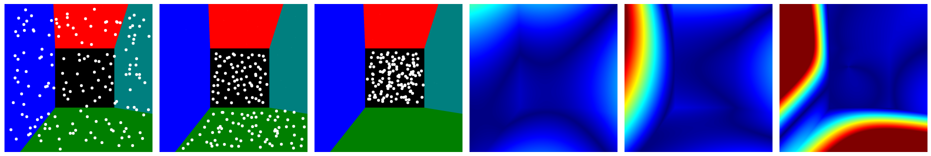

Fig. 1. Left three images: 100 sampled 2D points (white dots) drawn from different combinations of scene planes. Right three images: Associated epipolar error for the left images given corresponding points in a second (not shown) view. Here, epipolar error is calculated using the pair-wise geometry obtained via five-point RANSAC. Blue pixels indicate no epipolar error for the point in the image, while red pixels indicate a distance from the estimated epipolar line of 10 or more pixels. Even though the general camera pose is incorrect for the two- and one-plane examples, epipolar error is low within the sampled planes. However, large error (> 10 pixels) exists on the unsampled planes, especially in the single-plane case.

pose from such data in a calibrated setting, the five-point algorithm is usually embedded in a RANSAC [6] framework [21]. Here, a minimal number of random correspondences are sampled from the set of potential matches, the geometry estimation algorithm is run, and some subset of the complete set of correspondences is used to select the best of the ten resulting hypotheses. This sampling is repeated for a sufficient number of RANSAC iterations, and the solution with the strongest support, determined by the number of correspondences or a robust metric such as MSAC [27], is chosen as the optimal epipolar geometry.

In RANSAC, the distribution of valid point correspondences plays a strong role in determining the finally estimated two-view transformation. Take, for example, Fig. 1, which shows epipolar error maps for three different correspondence set scenarios. Here, we have added a small amount of noise and outliers to the point correspondence sets and have limited the available points to lie on five, two, and one scene plane. We run five-point RANSAC on these point sets to obtain a best-fitting epipolar geometry based on the criterion of maximal inlier support. It is clear that when all planes in the scene are equally sampled, the epipolar error is low throughout the image. When only one or two scene planes are available, however, the estimated pair-wise geometries have good support on the sampled scene planes, but they completely fail to accurately characterize unsampled scene planes – and the underlying relative pose is necessarily wrong. This short experiment serves to demonstrate that, even if the inlier support for an estimated calibrated two-view geometry "looks" correct in highly sampled regions, the underlying transformation that is estimated may actually very poorly fit unsampled scene planes. The motivation for our approach is to leverage point correspondences on other scene planes, when at least a small number are available, to solve this problem.

Fig. 2 demonstrates the plane-sampling problem for a real-world image pair. Especially for community photo-collections, it is common that points on buildings (which have more favorable views) are extremely well-matched using common feature descriptors like SIFT. Point correspondences on the ground, how-

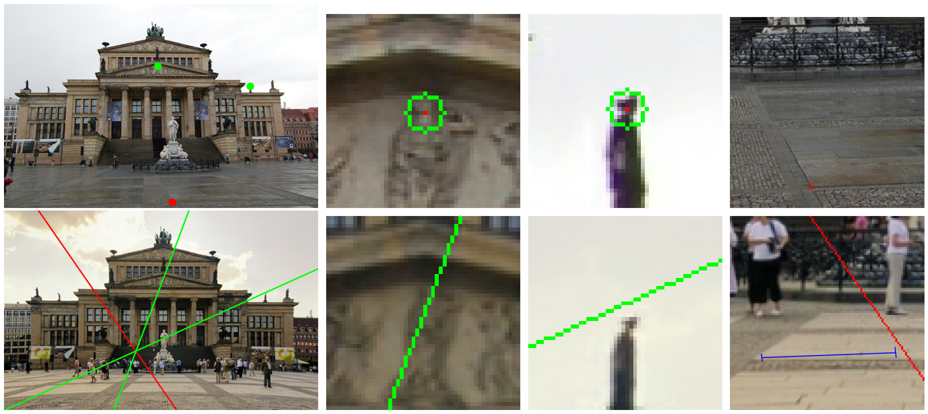

Fig. 2. Epipolar geometry estimated using a state-of-the-art SfM pipeline [23] for an example image pair. The upper left image illustrates a point on the ground area (red circle) and two non-ground points (green circles). On the lower left, the three corresponding epipolar lines (red and green) are illustrated on the second image. The zoomed images for the three points (right three columns) show that the epipolar line has low error for the two green points. The epipolar line for the point on the ground plane (right column), however, is inaccurate by more than 30 pixels (blue line).

ever, are often embarrassingly underrepresented. Accordingly, RANSAC very often fails to correctly find inlier point pairs on the ground surface, and the estimated epipolar geometry is globally inaccurate. For the interested reader, we provide in the supplementary material a systematic evaluation of the error behavior of the five-point algorithm in a RANSAC framework.

4 Proposed Method

Given our analysis, it is clear that the preservation of ground points during feature verification is a necessary prerequisite for an accurate estimation of the camera geometry from image pairs. In principle, a verification within the RANSAC procedure is possible without initial correspondence determination. To do so, for all solutions the complete image could be used for verification, *e.g.*, by dense matching or extended correspondence search for multiple pixels. Unfortunately, in a RANSAC procedure this would tremendously increase the runtime.

To efficiently obtain correct solutions for the two-view geometry, we propose to employ two distinct matching pipelines, one for the stably detected features of an image and one for features on the non-dominant planes, chiefly the ground plane. The second pipeline gives us a way to maintain efficiency in the estimation. We propose to perform the verification embedded in a RANSAC framework two times: Once for the separated ground areas and a second time for the complete images. This results in a sufficient number of matched ground points, which is essential for achieving accurate SfM including ground scene parts.

We start with a calibrated pair of images and detect SIFT features for both images. Matching is performed using standard SIFT matching with nearest neighbor search under the L_2 norm and a final ratio-based filtering [17]. Successfully matched points are then verified by means of the five-point algorithm within a RANSAC framework. This can lead to missing ground points and, hence, to an inaccurate registration of ground areas. Fig. 3 shows an example image pair with verified feature points.

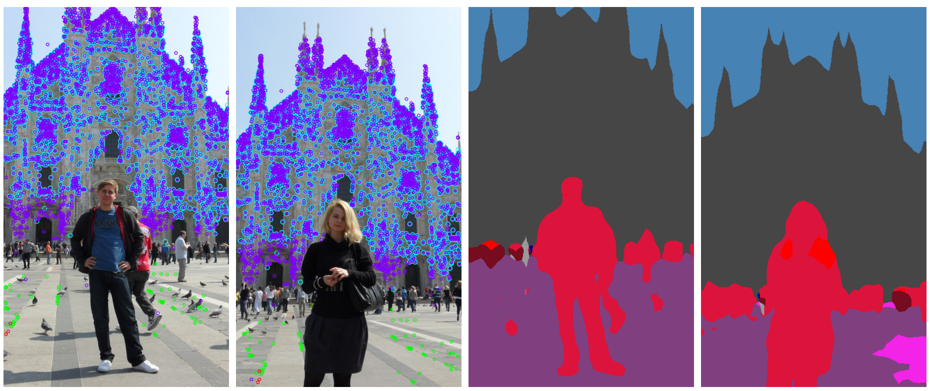

Fig. 3. Left two images: Example image pair marked with feature points. Right two images: semantic labels, including ground (purple) and building (grey). The purple points on the image pairs represent the feature points verified by the five-point algorithm in a standard RANSAC framework. On the ground plane, no (correct) verified point correspondences were found. The green and red points are the inliers from the first stage of our pipeline, which employs a homography for ground points. Turquoise and green points are the final features verified by the second RANSAC procedure.

To avoid behavior where the verification optimizes only specific areas, we propose to match the features for separate classes. To this end, one needs to identify specific semantic regions, such as the ground, in an image. Fortunately, the availability of plenty of labeled data for urban scenes [4] and the recent progress in classification by trained deep CNNs allows for a stable classification of multiple classes including building and ground [30]. We use the latter method to initially label our input images semantically (Fig. 3).

In our matching pipeline, we separate features from the image pairs into two classes: (1) road and sidewalk, and (2) building and wall [4, 30]. For our experiments, we ignore other labels, and we ignore correspondences if their classes do not agree. Our first matching and verification stage only makes use of the feature points assigned to the ground. Unfortunately, the matching of ground features usually leads to a high outlier rate. In typical image sets from, *e.g.*, community photo collections, the ground is only captured at an oblique angle. This is typical when using images acquired with a hand-held camera. In addition to perspective deformations, the ground is challenging to match because of many repetitions

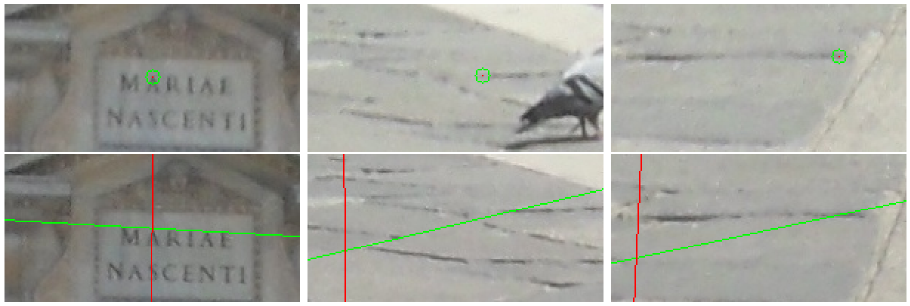

Fig. 4. Top row: Three example points from the image pair shown in Fig. 3. Bottom row: corresponding epipolar lines. The red line represents the epipolar line from standard five-point verification of SIFT feature matching by means of RANSAC, and the green line the epipolar line from our method. While both methods are highly accurate on the building (left column), only our method produces accurate epipolar geometry for the ground. The epipolar error for the standard verification is up to 100 pixels.

and a lack of unique structures. Thus, to increase the inlier ratio of ground point correspondences, we process the ground separately first by a ground-to-ground homography assuming a mostly planar ground.

To detect and match features on obliquely viewed surfaces, methods exist based on post-processing of descriptors [14] or a preceding affine transformation of the areas. For the reconstruction of highly slanted areas, complex pipelines allow for the generation of a small number of highly stable matches [19, 20]. Unfortunately, the methods from [19, 20] are not runtime-efficient. On the other hand, the matching has to be performed on a fraction of the complete image. We empirically found that also using standard SIFT allows ground matching for image pairs with small baselines. If the configuration needs complex matching methods, we use the implementation of [20]. Fig. 3 shows the resulting SIFT features after homography-based verification (green and red points).

For a joint representation of the point sets for road and sidewalk as well as building and wall, we use a second verification procedure. The input is the two sets of feature points: (1) The set of unverified SIFT features labeled as building or wall, and (2) the verified ground matches from the first procedure. For every RANSAC iteration, five points are randomly chosen from the first set. The up-to-ten solutions for the essential matrices are evaluated against all other points from this set. In our pipeline, the solution is additionally evaluated against the second set containing the stable ground points. Hence, we have two sets of inliers: ground inliers $\#I_g$ and building inliers $\#I_b$. To estimate the optimal solution from the set of hypotheses, we compare them against the current best solution with inlier sets $\{I_g^*, I_b^*\}$ by defining the quality as: $q_i = \frac{\#I_g^i}{\#I_g^*}\frac{\#I_b^i}{\#I_b^*}$. Hence, the relative ratio of verified ground and building points has to be maximal. The maximum quality over all hypotheses i is chosen for the optimal essential matrix. In Fig. 3, the finally verified points from the second procedure are presented in the form of green ground points and turquoise building points. Fig. 4 demonstrates the improvement achieved with our method. It is immediately apparent that for the ground area, the epipolar lines for the first verification have an error of tens of pixels, while our complete method achieves a high accuracy. Moreover, on the building, our method yields a similar quality as the essential matrix derived by the first verification.

5 Experiments

We have evaluated our method on challenging real-world image pairs from community photo-collections to demonstrate the improvement for the accurate estimation of the relative geometry. For the calibration (camera intrinsics), we make use of the EXIF information in the images.

First, the choice of feature detection can be crucial. We use SIFT features for the initial estimation of the matching points for ground and building, but as previously mentioned, these features sometimes have poor performance when matching ground points from wide-baseline pairs. Using SIFT, ground points are typically well-matched for similar small-baseline images (see Fig. 5). However, when even mild changes in appearance or perspective occur, or if the ground is weakly textured, the ground can prove challenging to match. In such cases, we use MODS [19, 20] for ground-point matching (Fig. 6). While other feature detection methods have been published recently [8, 29], we empirically found that none of them surpasses SIFT in producing reliable matches on obliquely viewed surfaces on our test images.

We use our proposed two-stage RANSAC procedure to solve for the epipolar geometry of the image pairs. More precisely, we use the extension LO-RANSAC [2], as our experience shows it generates better results for both the original veri-

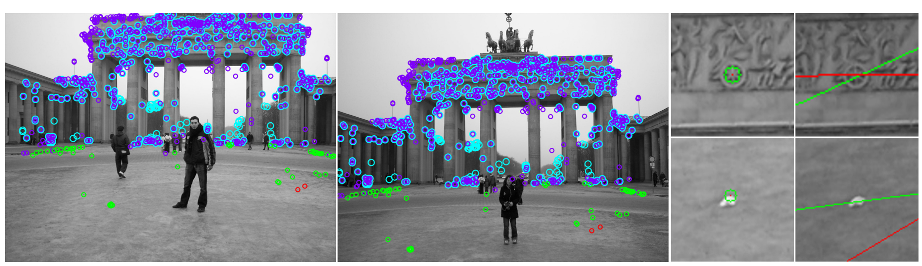

Fig. 5. Results of our method on an image pair using SIFT features. Left: Images with original inlier points shown in purple. The green and red points are the inliers from the first stage of our pipeline. Turquoise and green points are the final features verified by the second RANSAC procedure. Right: The zoomed images show a ground and a non-ground match. While the non-ground matches are correct for both methods, the ground points were only aligned correctly with our method (green epipolar line). The red epipolar line, representing the five-point essential matrix, shows an error of tens of pixels for the ground.

Fig. 6. Four examples for optimized essential matrices from image pairs. Because of challenging texture, perspective, resolution and lighting conditions in these examples MODS is used for ground-point matching. Verified features are color-coded as in Fig. 5. The zoomed images show examples for ground and non-ground matches. While the non-ground matches are correct for both methods, the ground points were only related correctly with our method (green epipolar line). The red epipolar line, representing the five-point essential matrix, shows an error of tens of pixels for ground images.

fication and our method. RANSAC is parameterized as follows: (1) a maximum error of 2 pixels, (2) a confidence value of 99.99%, and (3) a minimum inlier ration of 20%. Because the matching of the ground produces a higher outlier ratio, in the homography-based procedure a maximum error of 4 pixels is used. The higher maximum error takes into account that, often, the ground cannot be exactly described by a plane. Also, in the second procedure, the ground points are verified with double the pixel accuracy threshold of the building points; this accounts for the fact that the perspective transformation of ground points usually results in lower-quality feature localization and description.

Figs. 5 and 6 show several results for our method in comparison to the standard five-point-based estimation employing the complete set of SIFT features. Furthermore, in the supplementary material more results are provided. Our method allows the matching of ground points and the derivation of a highly accurate essential matrix. In situations where the traditional method only produces a sufficient accuracy for areas representing nearly fronto-parallel surface planes, our method produces an accuracy within a few pixels for the entire image. We assume a view without excessive perspective deformation of the building, which is given for typical images showing building and ground in similar proportion as in this case the building is generally captured from a larger distance or with a camera facing the facade in a fronto parallel manner. Because the camera calibration is not optimal, for some examples our verified correspondences do not cover the complete scene. The upper row in Fig. 6 gives two examples with several missing correspondences. Nonetheless, it is apparent that the distance to the epipolar line is within a couple of pixels, whereas insufficient consideration of ground features can lead to pixel errors of up to 100 pixels (Fig. 6). The latter would lead to a completely false estimation of the relative camera pose, prohibiting the image pair from contributing to an accurate 3D reconstruction (see Fig. 2).

6 Conclusion

In this paper, we revisited the five-point algorithm and provided evidence of its potential shortcomings for scene configurations frequently arising in image-based 3D reconstruction. In these configurations, point correspondences on obliquely viewed surfaces — particularly the ground — are largely missed, and even though the resulting essential matrices describe parts of the image well, the representation of the entire scene is strongly distorted in regions with low support. In a RANSAC framework, the five-point algorithm offers a multitude of hypotheses that are verified using the available correspondences. Correspondences situated only on a single fronto-parallel plane, in particular, lead to a poorly estimated essential matrix; this is common in 3D urban modeling, where correspondences are mostly captured on facades, but missing on the ground plane.

Our solution allows the preservation of ground correspondences in addition to those on buildings in urban scenes. To this end, we leverage an existing classification method to semantically segment source images into ground and non-ground regions. Our semantic-based hypothesis scoring approach makes use of these labelings to ensure that the undersampled ground correspondences are still accurately captured during RANSAC-based two-view geometry estimation. Results on a large variety of scenes demonstrate the ability of our approach to successfully maintain dominant-plane correspondences while additionally recovering ground correspondences. In the future, we look to expand the use of semantic labels to other aspects of two-view geometry estimation, including fundamental matrix estimation and related robust statistical measures.

References

1. Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building rome in a day. Commun. ACM 54(10), 105–112 (2011)
2. Chum, O., Matas, J., Kittler, J.: Locally optimized ransac. In: DAGM (2003)
3. Chum, O., Werner, T., Matas, J.: Two-view geometry estimation unaffected by a dominant plane. In: CVPR. vol. 1, pp. 772–779 (2005)
4. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
5. Faugeras, O.D., Maybank, S.J.: Motion from point matches: Multiplicity of solutions. IJCV 4(3), 225–246 (1990), http://dx.doi.org/10.1007/BF00054997
6. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24(6), 381–395 (1981)
7. Frahm, J.M., Georgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.H., Dunn, E., Clipp, B., Lazebnik, S., Pollefeys, M.: Building Rome on a cloudless day. In: ECCV (2010)
8. Han, X., Leung, T., Jia, Y., Sukthankar, R., Berg, A.C.: Matchnet: Unifying feature and metric learning for patch-based matching. In: CVPR (2015)
9. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, ISBN: 0521540518, second edn. (2004)
10. Hartley, R., Kahl, F.: Optimal algorithms in multiview geometry. In: ACCV (2007)
11. Hartley, R.I.: In defense of the eight-point algorithm. PAMI 19(6), 580–593 (1997)
12. Hartley, R.I., Kahl, F.: Global optimization through searching rotation space and optimal estimation of the essential matrix. In: ICCV (2007)
13. Heinly, J., Schönberger, J.L., Dunn, E., Frahm, J.M.: Reconstructing the World* in Six Days *(As Captured by the Yahoo 100 Million Image Dataset). In: CVPR (2015)
14. Ke, Y., Sukthankar, R.: PCA-SIFT: A more distinctive representation for local image descriptors. In: CVPR (2004)
15. Kushnir, M., Shimshoni, I.: Epipolar geometry estimation for urban scenes with repetitive structures. PAMI 36(12), 2381–2395 (2014)
16. Li, H., Hartley, R.: Five-point motion estimation made easy. In: ICPR (2006)
17. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)
18. Mayer, H., Bartelsen, J., Hirschmüller, H., Kuhn, A.: Dense 3D reconstruction from wide baseline image sets. In: 15th International Workshop on Theoretical Foundations of Computer Vision (2011)
19. Mishkin, D., Matas, J., Perdoch, M.: MODS: fast and robust method for two-view matching. CoRR abs/1503.02619 (2015), http://arxiv.org/abs/1503.02619
20. Mishkin, D., Matas, J., Perdoch, M., Lenc, K.: Wxbs: Wide baseline stereo generalizations. CoRR abs/1504.06603 (2015), http://arxiv.org/abs/1504.06603
21. Nistér, D.: An efficient solution to the five-point relative pose problem. PAMI 26(6), 756–777 (2004)
22. Rodehorst, V., Heinrichs, M., Hellwich, O.: Evaluation of relative pose estimation methods for multi-camera setups. International Archives of Photogrammetry and Remote Sensing (ISPRS) pp. 135–140 (2008)
23. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016)

24. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the world from internet photo collections. IJCV 80(2), 189–210 (2008)
25. Stewénius, H., Engels, C., Nistér, D.: Recent developments on direct relative orientation. ISPRS Journal of Photogrammetry and Remote Sensing 60, 284–294 (2006), http://dx.doi.org/10.1016/j.isprsjprs.2006.03.005
26. Torr, P.H.: An assessment of information criteria for motion model selection. In: Computer Vision and Pattern Recognition (CVPR). pp. 47–52 (1997)
27. Torr, P.H., Zisserman, A.: Mlesac: A new robust estimator with application to estimating image geometry. Computer Vision and Image Understanding 78(1), 138–156 (2000)
28. Wu, C.: Towards linear-time incremental structure from motion. In: 3DV. pp. 127–134 (2013)
29. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: LIFT: Learned invariant feature transform. In: ECCV (2016)
30. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: ICLR (2016)
31. Zach, C., Klopschitz, M., Pollefeys, M.: Disambiguating visual relations using loop constraints. In: Computer Vision and Pattern Recognition (CVPR). pp. 1426–1433 (2010)