Structure from Motion Estimation with Positional Cues

Linus Svärm and Magnus Oskarsson

Centre for Mathematical Sciences, Lund University, Sweden
{linus,magnuso}@maths.lth.se

Abstract. We present a system for structure from motion estimation using additional positioning data such as GPS data. The system incorporates the additional data in the outlier detection, the initial estimates and the final bundle adjustment. The initial solution is based on a novel objective function which is solved using convex optimization. This objective function is also used for outlier detection and removal. The initial solution is then refined based on a novel near L_2 minimization of the reprojection error using convex optimization methods. We present results on synthetic and real data, that shows the robustness, accuracy and speed of the proposed method.

1 Introduction

A classic problem in computer vision is how to estimate both the camera motion and the structure of a scene given only image data of said scene. This is known as the "Structure from motion" or SfM problem. Today, there exist systems that take thousands of images as input and solve the SfM problem in an automatic way, see e.g. [1, 2].

A big problem for many of these systems is the occurrence of outliers in the data, i.e. points that fit the initial models but do not fit the final models. These outliers cause major problems for the non-linear optimization methods, and can often result in local minima. Another problem is the speed of convergence of the different parts of the reconstruction pipeline when the amount of input data and estimated model parameters grows very large.

To remedy this, new approaches using convex optimization have been introduced in the computer vision community over the last years, see e.g. [3–6].

Today cameras are ubiquitous and image data is readily available. In addition to the pure image data there is today in many cases more information available both about the camera position and the structure of the scene. This could be e.g. GPS data, user geo-tagged images or depth information of the scene. The additional information is often complementary in nature to the pure image information. We believe that this information should be used in conjunction with the image data in order to *simplify the estimation problem, make it more robust and make it faster*. In order to fuse data from different modalities in a consistent way, it is important to work with meaningful objective functions in the optimization.

J.-K. Kämäräinen and M. Koskela (Eds.): SCIA 2013, LNCS 7944, pp. 522–532, 2013.

Much of the work that has been done incorporating both visual and positional data are set in a real-time framework, in a SLAM setting see e.g. [7]. Incorporating positional information in systems such as these is usually done in the final bundle adjustment. A number of contributions exist that incorporate e.g. GPS information into the final optimization, see e.g. [8–12] or odometric data [13–15]. If the initial estimates are not good enough this could lead to problems with local minima, which we show in the experimental section. We have found that incorporating additional known positional cues during the whole estimation process can make a large difference on the final reconstruction.

In this paper we will focus on a batch setting, where we try to solve the whole SfM problem with all the data available. We will show how additional positional data can be incorporated in the complete structure from motion estimation framework, and in this way both make the solutions more robust and more accurate as well as in some cases even speed up convergence. In order to do this we formulate a number of error measures that incorporate both image and positional data. We show, using methods from convex optimization, how these minimization problems can be solved in an efficient way. We have focused on estimating the translation of the cameras and 3D points and assume that the orientation of the cameras is known. There are a number of efficient methods for estimating consistent rotations between cameras, see [16–18], and any of these can be used in conjunction with our system.

Our key contributions in this paper are:

- A system for structure and motion estimation based on image data and additional positional data. The positional cues can be used in all steps of the estimation.
- An approximate L_2-norm formulation of the reprojection error. The goal function can be solved globally optimal using a novel method based on convex optimization.

2 Preliminaries

This section presents some theory and concepts which will be needed in the following sections. Additional information and proofs on convex optimization and SOCP problems can be found in [19].

Given a *quasiconvex* objective function f and convex constraints f_i, we can decide whether the optimal function value f^* is larger or smaller than some $\gamma \in \mathbb{R}$, by solving the convex feasibility problem

$$\begin{aligned} &\text{find} && x && (1)\\ &\text{subject to} && f(x) \leq \gamma && (2)\\ & && f_i(x) \leq b_i, i = 1, \ldots, m. && (3) \end{aligned}$$

These kinds of problems can be solved with great efficiency using interior point methods. As a consequence, quasiconvex problems can be minimized using bisection. For a particular class of convex optimization problems, the constraints are of the form $\|A_i x + b_i\| \leq c_i^T x + d_i$. These constraints describe second order cones, and the corresponding problems are called *second order cone programs* (SOCP).

Given a set of image points, seen in a set of images, we would like to recover the scene structure and the relative motion between the cameras. If the measurement errors are assumed to follow the normal distribution, the statistically optimal solution is given by minimizing the L_2 norm of the reprojection errors. Unfortunately, this is in general a very difficult function to minimize. Only local methods exist, which cannot guarantee global optimality. One way of handling this is to instead minimize the L_∞ norm of the error, see [3]. Using the L_∞ norm instead of the L_2 norm, makes it possible to find the global minimum.

Let u be an image point, U the corresponding estimated 3D point, and R and C orientation and center of the camera, respectively, so that with no noise we have:

$$\lambda \begin{bmatrix} u \\ 1 \end{bmatrix} = \begin{bmatrix} R & -RC \end{bmatrix} \begin{bmatrix} U \\ 1 \end{bmatrix}. \tag{4}$$

If

$$R = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix},$$

then we can write the reprojection error as

$$e = \frac{\left\|\left(r_1\left(U - C\right), r_2\left(U - C\right)\right) - \lambda u\right\|_2}{\lambda}, \tag{5}$$

where $\lambda = r_3\left(U - C\right)$ is the (unknown) depth of the point U. We consider the sublevel sets by setting $e \leq \epsilon$ and multiplying both sides by λ,

$$\left\|(r_1(U - C), r_2(U - C)) - \lambda u\right\|_2 \leq \epsilon\lambda. \tag{6}$$

When u, R and ϵ are known, this is a second-order cone. That means that the reprojection error is a quasiconvex function in U, C and λ. If there are many points and cameras, then the maximum reprojection error is still a quasiconvex function and can be minimized using bisection.

From now on, we assume that the rotational part of each camera is determined in advance. This is a reasonable assumption, and has become common practice in recent approaches. Methods for robust estimation of camera rotations are presented in [16–18, 20].

3 Framework

In this section the different steps of our system will be explained, including outlier removal, approximate L_2 optimization and incorporation of positional cues.

3.1 Outlier Removal

As we saw in the previous section, the structure and motion problem can be solved optimally in L_∞ norm by solving a sequence of convex feasibility problems. However, this is under the assumption that we have correct matchings between corresponding feature points in different images. This is rarely the case in real world scenarios. Thus, gross incorrect point matches, i.e. outliers in the data, need to be handled. We will here present a minimization scheme that can be used to remove outliers in a way similar to [21]. As suggested in [22], the cone constraints can be made more flexible using auxiliary variables. Equation (6) becomes

$$\|(r_1(U-C), r_2(U-C)) - \lambda u\|_2 \leq \epsilon\lambda + s, \tag{7}$$

where there is one $s \geq 0$ for every image point. This allows the reprojection error to become larger than the prescribed threshold ϵ. Ideally, we would like to minimize the number of active (i.e. non-zero) $s(i)$. This is however a very difficult problem, so following [22] we settle for the L_1 relaxation and minimize $\sum_i s(i)$ subject to $s(i) \geq 0$ for all i. When this minimization is complete, all outliers can be purged from our problem by removing all image points $u(i)$ for which $s(i) > 0$. Thus, solving one convex optimization problem, a solution can be found guaranteeing that the maximum reprojection error is smaller than ϵ.

3.2 Approximate L_2

As we have seen, using a bisection algorithm, we can find the optimal solution to the L_∞ problem. As discussed in section 2, what we really would like to minimize is the L_2 norm of the reprojection errors. This is asking to much. However we can formulate an approximation to the L_2 norm, as a SOCP. To see how, we consider again equation (5). We would like to solve:

$$\text{minimize}_i \, e^2(i). \tag{8}$$

Thus, we start by looking at e^2:

$$e^2 = \frac{((r_1 - u_1 r_3)(U-C))^2 + ((r_2 - u_2 r_3)(U-C))^2}{\lambda^2}, \tag{9}$$

where u_i is the ith component of u. Imagine for a moment the square in the denominator removed. Let us call this function g. To see that g is a convex function, recall that a function is convex if its epigraph is a convex set. The epigraph of g is given by $\{(x,t)|t \geq g(x))\}$. This can be written as

$$\left\{(x,t)|\frac{1}{2}(t + c^T x) \geq \left\|\left(a^T x, b^T x, \frac{1}{2}(t - c^T x))\right)\right\|_2\right\}, \tag{10}$$

which is a second order convex cone. We can write the minimization of g as

$$\begin{aligned} &\text{minimize} \quad t && (11)\\ &\text{subject to} \quad t \geq g(x). && (12)\end{aligned}$$

By using (10), we see that we can rewrite the constraint as a second order cone. Hence, we can formulate the minimization problem of g as the second order cone program

$$\text{minimize} \quad t \tag{13}$$

$$\text{subject to} \quad \|(h_1, h_2, h_3))\| \leq \frac{1}{2}(t + \lambda), \tag{14}$$

where, to keep notation compact, $h_1 = (r_1 - u_1 r_3)(U - C)$, $h_2 = (r_2 - u_2 r_3)(U - C)$, $h_3 = \frac{1}{2}(t - \lambda)$. We can extract depth estimates $\hat{\lambda}$ from the outlier removal step. Given these, we can use the above reasoning to approximate (9) by

$$\hat{e}^2 = \frac{h_1^2 + h_2^2}{\hat{\lambda}\lambda}. \tag{15}$$

To conclude our work so far, we summarize the preceeding sections in an algorithm for computing structure and motion using an approximate L_2 norm.

Algorithm 1. *(Approximate L_2 SfM)*

1. *Perform outlier removal as suggested in section (3.1), by solving*

$$\text{minimize}_i \; s(i),$$

subject to the constraints given by equation (7).

2. *Use the solution obtained during the outlier removal step to get depth estimates $\hat{\lambda}(i)$. Solve (8) using the approximation in (15).*

In the above parametrization, there is an ambiguity in the choice of coordinate system. Structure and motion is determined up to unknown scaling and translation. The approximate L_2 norm presented has a slight bias toward scaling down the solution. This is a result of using the approximate depth. To overcome this, and to be able to handle positional cues, we fix the scale by enforcing all depths to be larger than, or equal to, 1.

3.3 Incorporating other Sensors

In this section we show how to incorporate additional positional cues. Probably the most readily available measurements, besides the image itself, is GPS-data. We would like to be able to incorporate such information in our framework.

Let $\hat{C}(j)$ be a position measurement for camera j. If the measurement error is assumed to follow the normal distribution, the statistically optimal solution is to minimize the L_2 norm of measurement errors. Since the scale of the reconstruction is already fixed, as described above, we need to decouple scale from position measurements. We introduce a variable ς and write the L_2 norm of the measurement error as $\left\|\varsigma\hat{C}(j) - C(j)\right\|_2$.

Following the idea we used in section 3.2, we see that it is possible to formulate minimization of the scale invariant L_2 norm as the convex problem

$$\text{minimize} \quad \sum_i s(i) \tag{16}$$

$$\text{subject to} \quad \|P\|_2 \leq \frac{1}{2}(s(i)+1), \tag{17}$$

where $P = (\varsigma\hat{c}_1(i) - c_1(i), \varsigma\hat{c}_2(i) - c_2(i), \varsigma\hat{c}_3(i) - c_3(i), \frac{1}{2}(s(i)-1))$, and $c_i(j)$ is the ith coordinate of camera j.

Under the assumption that the different types of errors are independent, it is statistically optimal to weigh the errors by their variance, see the classic paper [23]. The modified objective function for the second optimization step becomes

$$\text{minimize} \quad \sigma_e^2 \sum_i t(i) + \sigma_{pos}^2 \sum_j s(j), \tag{18}$$

subject to constraints (14) and (17).

By taking camera measurements into consideration in the initial outlier removal step, the set of feasible solutions is pruned. Hence, the risk that an outlier fits into the solution is decreased. In the initial step, we optimize the L_1 relaxation of the number of outliers. A point should preferably be either an inlier or an outlier. This motivates us to use hard constraints. Thus, we decide on a threshold and require all cameras with measurements to be placed within the threshold distance from the measurements. Using the same notation as above the constraints are of the form

$$\left\|\varsigma\hat{C}(i) - C(i)\right\|_2 \leq \omega, \tag{19}$$

where ω is the distance threshold.

We will not go through the details here, but it is more or less straight-forward to incorporate any additional positional cues on scene structure points that are available.

Algorithm 2. *(SfM with Positional Cues)*

1. *Outlier removal and depth estimation. Solve*

$$\text{minimize} \quad \sum_i s(i) \tag{20}$$

   subject to (7), (19), and $\lambda(i) \geq 1$
2. *Use the solution obtained during the outlier removal step to get depth estimates $\hat{\lambda}(i)$. Solve*

$$\text{minimize} \quad \sigma_e^2 \sum_i t(i) + \sigma_{pos}^2 \sum_j s(j), \tag{21}$$

$$\text{subject to} \quad \|(h_1, h_2, h_3))\| \leq \frac{1}{2}(t(i) + \lambda)), \tag{22}$$

$$\|P\|_2 \leq \frac{1}{2}(s(j) + 1), \tag{23}$$

$$\lambda(i) \geq 1. \tag{24}$$

3. Do bundle adjustment on L_2 norm.

4 Experiments

In this section we test our system in a number of experiments using both synthetic and real data. The software SeDuMi is used to solve all convex problems.

4.1 Experiments on Simulated Data

We use synthetic data in order to be able to compare with ground truth camera positions. Here we present two such scenarios. In the first one, an imaginary street was placed along a circular arc. Along the sides of the street 3D points were put on facades. Equidistant cameras were placed along the street, seeing some of the points. Each point was registered on the cameras image plane with Gaussian distributed error (standard deviation 0.04). Further, each camera was annotated with position measurements, also with Gaussian errors to simulate GPS-data. In [8], the typical standard deviation of a consumer GPS-device is given as 2.34 m. This is what we use in our synthetic experiments. Finally, 10 percent of all point correspondences between cameras were mismatched, to simulate gross outliers. The second scenario was constructed in the same way, but with a different geometry with uniformly distributed GPS error (0 to 4 m) with all cameras in a spiral pattern. For each scenario, results with and without GPS data can be seen in figure 1 and table 1. Looking at the table, we see that

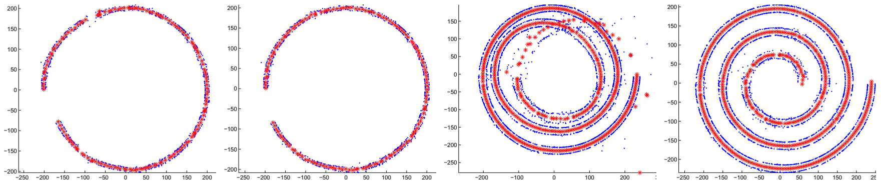

Fig. 1. Comparison between SfM estimation for synthetic city streets. Without using GPS cues (left circle and spiral) and using GPS cues (right circle and spiral). Red dots are camera positions, blue dots are the estimated structure.

Table 1. Results from the synthetic experiments. The standard deviation of the errors in the estimated 3D points and camera positions are given by σ_U and σ_C respectively. All measures are given in meters.

	Circle		Spiral	
	no gps	gps	no gps	gps
RMS	6.06	2.35	4.33	1.54
σ_U	3.32	1.13	1.57	0.92
σ_C	5.78	1.01	34.6	0.71

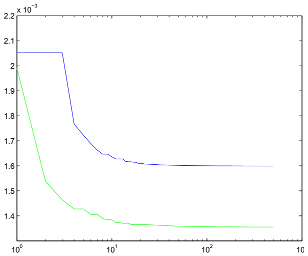

Fig. 2. Here the L_2 reprojection error is shown as a function of the number of iterations. The top blue curve shows the convergence without GPS information and the bottom green curve shows the convergence using the GPS information.

using GPS information, the standard deviation of the camera position error (i.e. distance between estimated camera position and ground truth) had a standard deviation σ_C of merely 1.01 and 0.71 for the two scenarios. We also get smaller errors in the estimated 3D points (standard deviation σ_U). This despite the simulated GPS-errors being much larger.

4.2 Experiments on Real Data

We have conducted experiments on a number of street-view images with additional positional data available for each frame. The positional data is assumed to have been acquired using GPS and inertial sensors, but we do not have apriori information about the errors in these measurements. The images are 360 degree panoramas that were rotationally registered in a common frame during acquisition. This means that the orientations of all cameras are known. In figure 3 reconstructions from two different datasets are shown. For each dataset we have reconstructions with and without the positional cues. The reconstructions on the left hand side are without positional cues. As can be readily seen there are a number of problems with the reconstructions without the positional cues that are remedied using them. Run times and number of points and cameras for the two setups are shown in table 2. In figure 2 the convergence of the optimization is compared with and without using the positional cues. Here the top curve shows the behavior without the positional cues. As can be seen the bottom green curve converges both faster and to a lower minimum. This is rather surprising

Table 2. Runtimes (in seconds) and number of points and cameras for the different experiments, and the different steps in Algorithm 2

Experiment	Nr of Cameras	Nr of points	step 1	step 2	Bundle Adjustment	step 1 with GPS	step 2 with GPS	BA with GPS
Circle	146	1082	8.85	2.91	7.74	7.71	3.42	5.42
Spiral	405	3422	24.64	8.77	32.86	13.75	9.19	33.59
City 2	332	21422	678.23	391.61	600.76	647.12	374.10	513.21
City 1	1330	79496	250.13	163.42	237.99	249.70	130.6	269.54

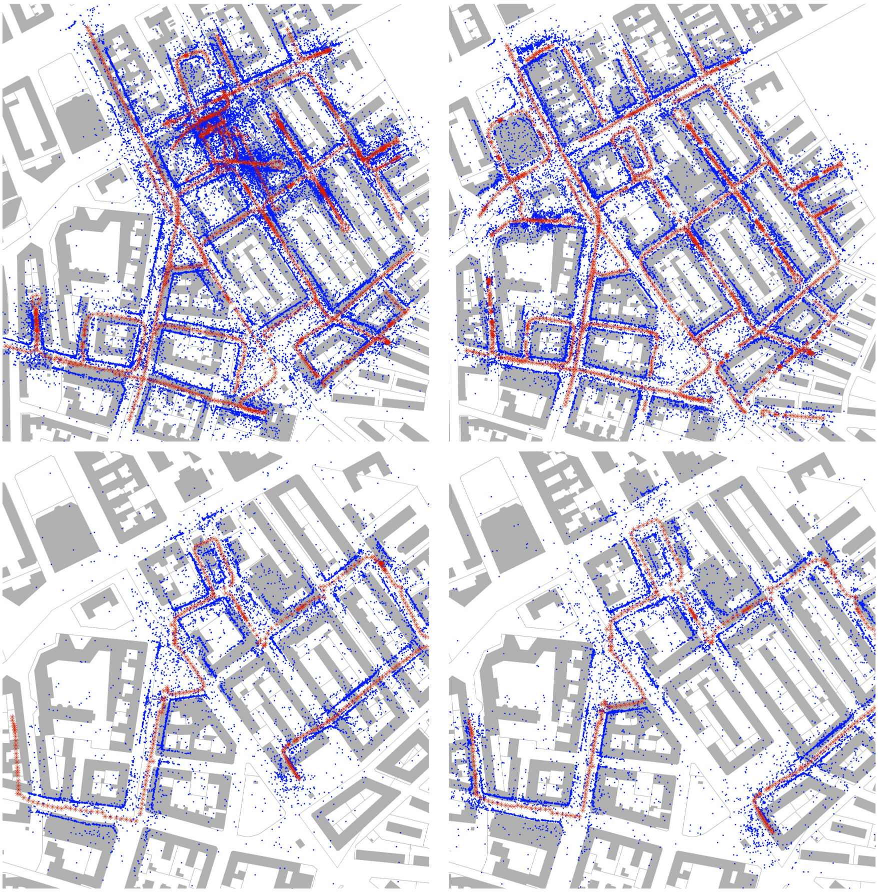

Fig. 3. Comparison between SfM estimations for a city section. Without using GPS cues (top left) and using GPS cues (top right). Both reconstructions are registered to a GIS model of the city section. Without using GPS information during step 1 and 2, the solution gets stuck in a local minimum. The second city section without using GPS cues (bottom left) and using GPS cues (bottom right). The solution without GPS looks decent, but suffers from drift.

since it is the L_2-norm of the reprojection errors that is shown. Incorporating the positional cues in the optimization will add terms to goal function. That we still reach a lower minimum means that we have found a better optimum and the blue curve has found a local optimum.

5 Conclusion

We have presented a complete system for doing structure from motion estimation. In this framework we have also shown how to incorporate additional positional cues such as e.g. GPS data into the optimization. The positional data is used throughout the optmization process, which includes an initial global optimization based on the L_∞-norm which is also used to root out outliers. This is then followed by an approximate formulation of the L_2-norm of the reprojection errors which is also solved globally. This is finally used in a bundle adjustment. The process has been shown to be robust and accurate in a number of experiments.

Acknowledgments. This work has been financed by the project VINST, grant no. RIT08-0043, funded by the Swedish Foundation for Strategic Research.

References

1. Agarwal, S., Snavely, N., Simon, I., Seitz, S.M., Szeliski, R.: Building rome in a day. In: Proc. 12th Int. Conf. on Computer Vision, Kyoto, Japan (2009)
2. Crandall, D., Owens, A., Snavely, N., Huttenlocher, D.: Discrete-continuous optimization for large-scale structure from motion. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2011)
3. Hartley, R., Schaffalitzky, F.: L_∞ minimization in geometric reconstruction problems. In: Proc. Conf. Computer Vision and Pattern Recognition, Washington DC, USA, pp. 504–509 (2004)
4. Kahl, F.: Multiple view geometry and the L_∞-norm. In: International Conference on Computer Vision, Beijing, China, pp. 1002–1009 (2005)
5. Ke, Q., Kanade, T.: Quasiconvex optimization for robust geometric reconstruction. In: International Conference on Computer Vision, Beijing, China, pp. 986–993 (2005)
6. Zach, C., Pollefeys, M.: Practical methods for convex multi-view reconstruction. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 354–367. Springer, Heidelberg (2010)
7. Nistér, D., Naroditsky, O., Bergen, J.: Visual odometry. In: Proc. Conf. Computer Vision and Pattern Recognition, Washington DC (2004)
8. Lhuillier, M.: Fusion of gps and structure-from-motion using constrained bundle adjustment. In: CVPR. IEEE Computer Society (2011)
9. Pylvänäinen, T., Fan, L., Lepetit, V.: Revisiting the pnp problem with a gps. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Kuno, Y., Wang, J., Wang, J.-X., Wang, J., Pajarola, R., Lindstrom, P., Hinkenjann, A., Encarnação, M.L., Silva, C.T., Coming, D. (eds.) ISVC 2009, Part I. LNCS, vol. 5875, pp. 819–830. Springer, Heidelberg (2009)

10. Hu, Z., Keiichi, U., Lu, H., Lamosa, F.: Fusion of vision, 3d gyro and gps for camera dynamic registration. In: Proc. International Conference on Pattern Recognition, Cambridge, UK (2004)
11. Strecha, C., Pylvanainen, T., Fua, P.: Dynamic and scalable large scale image reconstruction. In: Proc. Conf. Computer Vision and Pattern Recognition, Colorado Springs, USA (2010)
12. Lothe, P., Bourgeois, S., Dekeyser, F., Royer, E., Dhome, M.: Towards geographical referencing of monocular slam reconstruction using 3d city models:applications to real-time accurate vision based localization. In: CVPR. IEEE Computer Society (2009)
13. Michot, J., Bartoli, A., Gaspard, F.: Bi-objective bundle adjustment with application to multi-sensor slam. In: 3DPVT (2010)
14. Konolige, K., Agrawal, M., Sola, J.: Large scale visual odometry for rough terrain. In: Intl. Symp. on Robotics Research (1996)
15. Strelow, D., Singh, S.: Motion estimation from image and inertial measurements. Int. Journal of Robotics Research 23(12), 1157–1195 (2004)
16. Govindu, V.M.: Robustness in motion averaging. In: Narayanan, P.J., Nayar, S.K., Shum, H.-Y. (eds.) ACCV 2006. LNCS, vol. 3852, pp. 457–466. Springer, Heidelberg (2006)
17. Zach, C., Klopschitz, M., Pollefeys, M.: Disambiguating visual relations using loop constraints. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1426–1433 (June 2010)
18. Martinec, D., Pajdla, T.: Robust rotation and translation estimation in multiview reconstruction. In: CVPR. IEEE Computer Society (2007)
19. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004)
20. Hartley, R., Aftab, K., Trumpf, J.: L1 rotation averaging using the weiszfeld algorithm. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3041–3048 (June 2011)
21. Sim, K., Hartley, R.: Removing outliers using the L_∞-norm. In: Proc. Conf. Computer Vision and Pattern Recognition, New York City, USA, pp. 485–492 (2006)
22. Dalalyan, A., Keriven, R.: L1-penalized robust estimation for a class of inverse problems arising in multiview geometry. In: 23d Annual Conference on Neural Information Processing Systems, Vancouver, Canada (December 2009)
23. Aitken, A.: On least squares and linear combination of observations. Proc. R. Soc. Edinb. 55, 42–48 (1934)