Effective and Generic Structure from Motion using Angular Error

Maxime Lhuillier
LASMEA, UMR CNRS 6602, Université Blaise Pascal, 63177 Aubière, France
Maxime.Lhuillier@univ-bpclermont.fr, maxime.lhuillier.free.fr

Abstract

Generic camera modeling using raxels and associated methods was recently introduced in Computer Vision. The main advantage is the applicability for any camera model, which contrasts to the many specific methods designed for a single camera model. This paper introduces a bundle adjustment based on angular error for the generic structure from motion problem. Experiments include automatic, robust and optimal estimation of scene structure and camera motion from a long image sequence acquired by a hand-held non-central (calibrated) catadioptric camera.

1. Introduction

The automatic estimation of scene structure and camera motion from an image sequence (“Structure from Motion”, or **SfM**) acquired by a hand-held camera is a fundamental topic in Computer Vision. Both specific (camera model dependent) and generic SfM methods were proposed.

Specific SfM A large amount of works has been done during the last decades for many camera models including perspective, stereo rig, and omnidirectional (fish-eye and catadioptric) cameras. Many successful systems now exist for image sequences acquired by the perspective and stereo rig models [15, 2, 13]. In contrast to these, mainly two views SfM works have been conducted for both central [3, 9] and non-central [1, 11] omnidirectional cameras, except [10]. A central camera is such that all back-projected rays intersect a single point in space [5], called the camera center.

Generic SfM Generic SfM methods could be preferred to the many methods designed for a single camera model. An arbitrary imaging system may be modeled by a general model composed of a set of virtual sensing elements called raxels [6]: raxels are central cameras with a small part of the complete view field whose centers are spread in the loci of view points of the system. Given a raxel discretization of a calibrated system, the system motion can be recovered using generalizations of pose calculation [12] or essential matrix [14]. More recently [16], this matrix is estimated linearly using 17 correspondences and the resulting geometry mixing different cameras (pinhole, stereo, central fish-eye) is refined using two possible bundle adjustments. The first one minimizes a 3D error. Such 3D errors [16, 11] are defined between a ray and a point in 3D, and are not ideal since 3D magnitude orders may vary considerably between close and far away points from cameras (the minimization result is biased in disfavor of the closest 3D points). 3D errors were introduced since the projection function of some camera models (e.g. non-central catadioptric) are not explicit, and difficult to minimize using bundle adjustment. The second one minimizes the image errors (measured in pixels) of perspective cameras as raxels. It requires a segmentation of all camera rays in clusters, such that all rays in a cluster are approximated by the rays of a perspective camera to be estimated. According to [16], finding the raxel number and the set of rays approximated by each raxel is yet the subject of future research for non-trivial cases.

Contributions We introduce a bundle adjustment based on an angular error designed for generic SfM, which has many advantages over previous methods. First, statistical foundation and error propagations are possible; this contrasts to the 3D errors [16, 11] producing biased results in many cases. Second, the angular error does not require a segmentation of the set of all camera rays, which contrasts to the perspective image error minimized in [16]. Third, angular error calculations are really faster than the specific image calculations of the non-central catadioptric case.

We also experiment the angular bundle adjustment in a non-trivial context: the automatic, robust and optimal estimation of a long image sequence acquired by a non-central catadioptric camera. Other experiments including uncertainty, accuracy and piecewise planar modeling are given.

2. Angular Error

Notations Let $x_{i,j}$ be many matched image points detected in the i-th image and corresponding to the j-th 3D scene point to be estimated. Since the camera is calibrated, the corresponding ray of $x_{i,j}$ is known. This ray is the oriented line in 3D which goes across point $s_{i,j}$ with direction $\vec{d}_{i,j}$. Direction $\vec{d}_{i,j}$ is fixed in the i-th camera coordinate system, and point $s_{i,j}$ has one d.o.f. along the line in 3D.

We also introduce the orientation R_i (a rotation) and the origin t_i (a 3D vector) of the i-th camera coordinate system defined in the world coordinate system, and X_j the world coordinates of the j-th 3D point.

Ray Surface Choice Point $s_{i,j}$ should be fixed as a starting point of the ray, since our angular error and the SfM result will depend on it. Although it is natural to locate $s_{i,j}$ at the camera center for a central camera, the choice is not so obvious in all cases. A choice in a general context is proposed by [6]: $s_{i,j}$ is taken in the caustic surface which is the envelope of all possible camera rays. The caustic has two interesting properties [6]: $\vec{d}_{i,j}$ is tangent to the caustic at point $s_{i,j}$, and the caustic is the locus of points where incoming rays most converge. An other $s_{i,j}$ choice for catadioptric camera may be the reflexion point of the ray on the mirror, or the symmetry axis if any. Such choices are intuitive (no mathematical justification given), and we assume that the ray surface is given by calibration. Once a ray surface (the set of $s_{i,j}$) is chosen, the (i,j)-th ray is re-defined by the half line starting from point $s_{i,j}$ with direction $\vec{d}_{i,j}$.

Angular Error In the ideal case, the (i,j)-th ray $(s_{i,j}, \vec{d}_{i,j})$ goes across X_j using previous notations. This is not the case in practice, since the detected $x_{i,j}$ are corrupted by image noise. We measure the gap between the (i,j)-th ray and point X_j by the angular error $e_{i,j}$, defined by the angle between the direction $\vec{d}_{i,j}$ and the direction $\vec{D}_{i,j}$ of the half line starting from 3D point $s_{i,j}$ toward 3D point X_j. Now, our formalization for the structure from motion problem is the following: the problem solution is a set of parameters R_i, t_i, X_j such that the sum of squared $e_{i,j}$ is minimal. A such angular error has three advantages.

First, it does not involve the camera model knowledge during SfM computations: the model is only required once for each ray $(s_{i,j}, \vec{d}_{i,j})$ estimations, before all SfM calculations. Consequently, the angular error allows generic SfM. This contrasts to the usual image errors measured in pixels, which require the camera model for all SfM calculations. This is a clear advantage for the angular error when the image error is slow to calculate and derivate, as the non-central catadioptric case where image error has no closed form. Second, it provides a maximum likelihood estimation, as described at the end of Section 3. This contrasts to the 3D errors introduced to simplify the estimation for non-central models [11, 16]. Third, the segmentation problem mentioned in [16] does not occur here.

3. Generic Structure from Motion

Geometry Estimation from Image Sequence The geometry of a sequence is estimated using a hierarchical approach, which is well known for perspective cameras [8]: once the geometries of the two camera sub-sequences $1 \cdots \frac{n}{2}, \frac{n}{2}+1$ and $\frac{n}{2}, \frac{n}{2}+1, \cdots n$ are estimated, the latter is mapped in the coordinate system of the former thanks to the two common cameras $\frac{n}{2}, \frac{n}{2}+1$, and the resulting sequence $1 \cdots n$ is refined by a bundle adjustment. The angular bundle adjustment designed for generic SfM is described below.

The hierarchical approach requires the geometries of all consecutive image triples of the sequence. Many methods are possible to estimate the triple geometries given the rays $(s_{i,j}, \vec{d}_{i,j})$ defined in the camera coordinate systems, before the refinement by angular bundle adjustment. If the camera is central, old methods could be used: first, all pair geometries are estimated by the essential matrix [4], second all triple geometries are estimated by the pose calculation [7] of the third camera once matches in 3 views have been reconstructed by the two others. The principle is similar for any cameras (including non-central models) using more recent methods: the two and three views geometries are given by the generalizations of the essential matrix [14] and the 3-points pose method [12].

Effective Angular Bundle Adjustment The angular bundle adjustment improves the estimations of the j-th 3D point X_j, the orientation R_i and origin t_i of the i-th camera coordinate system in the world coordinates by the minimization of a score: the sum of squared angles $e_{i,j}$ between directions $\vec{d}_{i,j}$ and $\vec{D}_{i,j}$. Both directions are expressed in the i-th camera coordinate system. $s_{i,j}$ and $\vec{d}_{i,j}$ are fixed from 2D point $x_{i,j}$, the camera calibration and the ray surface choice. $\vec{D}_{i,j}$ is the direction of the half line starting from point $s_{i,j}$ toward point X_j. By identifying point X_j and its homogeneous world coordinate with the fourth one set to 1, we have $\vec{D}_{i,j} = \frac{R_i^\top \begin{pmatrix} I_3 & -t_i \end{pmatrix} X_j - s_{i,j}}{||R_i^\top \begin{pmatrix} I_3 & -t_i \end{pmatrix} X_j - s_{i,j}||}$.

At first glance, the global angular error to minimize might be $\sum_{i,j} e_{i,j}^2$ with $e_{i,j} = arcos(\vec{d}_{i,j}.\vec{D}_{i,j})$. However, this $e_{i,j}$ is not a $\mathcal{C}^1$ continuous function at the exact solution when $e_{i,j} = 0$ (proof in the appendix). Since a legal use of the Levenberg-Marquardt method (used by the bundle adjustment [8]) requires a $\mathcal{C}^2$ continuous function $e_{i,j}^2$ in a neighborhood of the exact solution, we prefer to revise the expression of $e_{i,j}$. A second try might be $e_{i,j} = f(\vec{d}_{i,j}.\vec{D}_{i,j})$ with f a decreasing $\mathcal{C}^2$ continuous function such that $f(1) = 0$, if we accept that $e_{i,j}$ is not an angle. Now, $e_{i,j}$ is $\mathcal{C}^2$ continuous and has a local extrema at the exact solution $\vec{d}_{i,j}.\vec{D}_{i,j} = 1$: the Jacobian of $e_{i,j}$ is zero here. The convergence rate of Levenberg-Marquardt might be reduced in this context. The final $e_{i,j}$ proposition has not these problems, and is defined as follows.

Let $R_{i,j}$ be a rotation such that $R_{i,j}\vec{d}_{i,j} = \vec{k}$ with $\vec{k} = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}^\top$, and $\pi(\begin{pmatrix} x & y & z \end{pmatrix}^\top) = \begin{pmatrix} \frac{x}{z} & \frac{y}{z} \end{pmatrix}^\top$, $\begin{pmatrix} \tilde{x}_{i,j} & \tilde{y}_{i,j} & \tilde{z}_{i,j} \end{pmatrix} = R_{i,j}\vec{D}_{i,j}$. Now, we propose to minimize $\sum_{i,j} ||e_{i,j}||^2$ with $e_{i,j} = \pi(R_{i,j}\vec{D}_{i,j})$. We have $||e_{i,j}||^2 = \frac{\tilde{x}_{i,j}^2 + \tilde{y}_{i,j}^2}{\tilde{z}_{i,j}^2} = tan^2(\vec{k}, R_{i,j}\vec{D}_{i,j}) = tan^2(\vec{d}_{i,j}, \vec{D}_{i,j})$.

This is an acceptable approximation of the squared angle between $\vec{d}_{i,j}$ and $\vec{D}_{i,j}$ since this angle is small for the inlier point near the solution. The final angular expression $\sum_{i,j} ||\pi(R_{i,j}(R_i^\top \, (I_3 \quad -t_i \,) \, X_j \; - \; s_{i,j}X_j^t))||^2$ is minimized, with X_j^t any value of the 4-th X_j homogeneous coordinate. We note that $e_{i,j}$ is a 2D vector, which is independent of the $R_{i,j}$ choice. Experiments confirm that the tangent approximation has the best convergence.

Link with Perspective-based Raxels Note that the angle approximation by 2D $e_{i,j}$ is the projection of X_j by a calibrated perspective camera with center $s_{i,j}$ and orientation $R_{i,j}$, all expressed in the i-th frame. In our case, the ray segmentation in clusters [16] is trivial: one ray=one cluster.

Maximum Likelihood Estimation and Uncertainty A maximum likelihood estimation for X_j, t_i, R_i is obtained by minimizing $\sum_{i,j} ||e_{i,j}||^2$, if we assume that the 2D-angular errors $e_{i,j}$ obeys independent and identical zero-mean Gaussian distributions. This Gaussian model is not tenable if $e_{i,j} = f(\vec{d}_{i,j}.\vec{D}_{i,j}) \geq 0$ is chosen at first glance. This perturbation of $e_{i,j}$ propagates to a Gaussian perturbation of the estimated parameters, such that the covariance matrix may be estimated [17, 8].

4. Experiments

A non-central catadioptric camera [10] is used. The mirror profile is a known four degree polynomial, with height and big radius of 3.3 and 3.7 cm. The pinhole camera center is located at 48 cm below the mirror apex on the mirror symmetry axis. The caustic profile size is about the half of that of the mirror. The view field is about 100 degrees.

Synthetic Experiments We compare the accuracies of reconstructions estimated by angular bundle adjustment for many ray surface choices. The ground truth reconstruction is composed of 1000 points well spread on a $2.6m \times 3.4m \times 2.45m$ box surface (indoor scene like), and 12 cameras in a (unclosed) ellipse trajectory with a 0.5/0.9m radii at the box center. First, image projections are corrupted by gaussian noise of $\sigma = 1$ pixel. Second, all resulting rays (oriented line in 3D) are estimated by mirror reflexion. Third, we choose a ray surface and the angular bundle adjustment is applied starting from ground truth. Finally, errors between the resulting (t_i, X_j) and ground truth (t_i^g, X_j^g) are given. The squared location error is $E_t^2 = \frac{1}{12}\sum_i ||S(t_i) - t_i^g||^2$ with S the similarity transformation minimizing E_t, and the squared reconstruction error is $E_x^2 = \frac{1}{1000}\sum_j ||S(X_j) - X_j^g||^2$.

Table 1 shows similar errors for many ray surface choices and the current box "big box", including the choice $\{0\}$ (central approximation). For this 3D scale and noise level, we note that the ray surface choice is not so important and that the central model is a good approximation of the non-central camera (angles do not change much if the 3D point depths are much bigger than the size of area where the ray surface is selected). The same experiments are redone with the same scene reduced by 10 in the 3 dimensions "small box". Only small differences are noted for our ray surface choices, except for the central errors which are the worst.

ray surface	$\{0\}$ (central)	mirror	axis	caustic
E_t/E_x (big)	0.109/1.905	0.106/1.719	0.108/1.719	0.107/1.713
E_t/E_x (small)	0.040/0.901	0.010/0.203	0.013/0.185	0.012/0.183

Table 1. E_t, E_x errors (cm) for many ray surface choices.

Automatic SfM for Real Sequences A SfM strategy suggested by the synthetic experiments is used: (1) approximate the camera with the central model ($\{0\}$ choice) and apply hierarchical SfM (2) upgrade the central reconstruction to the non-central one with a ray surface choice (e.g. mirror). The generic, angular error-based bundle adjustment is used in both steps. A match $x_{i,j}$ is considered as outlier if $||e_{i,j}|| > 0.04$ radians. Such two-step and non-central SfM were previously used in non-generic contexts [11, 10]. More details are given in [10] about the calibration estimation, the automatic omnidirectional image border detection, and the specific matching method combining Harris points and ZNCC correlation. SfM based on [14, 12] is not tried.

Real Sequence The House sequence is composed of 112 images. The user moves along a trajectory on the ground with the omnidirectional system mounted on a monopod, alternating a step forward and a shot.

A top view of the resulting reconstruction and a piecewise planar 3D model obtained from the reconstructed points are shown in Figure 1. 18263 points are automatically reconstructed with 93235 inlier reprojections, and the final RMS angular error is 0.0057 radians. The planes of the model are estimated from 3D points and manual delimitation of their contours in 4 selected images (one for each room), ignoring occluded, uniform or too complex objects.

Uncertainties Uncertainties might be given using a trivial gauge constraint [17]: $R_0 = I_3, t_0 = 0$. A 7-th scalar relation like $(t_{111})_x = 1$ would not be necessary to fix the 3D scale factor, since it is theoretically possible to estimate this scale with the non-central model. However, further experiments shown that this estimation is difficult in practice for both specific and generic bundle adjustments. For this reason, we only give the realistic central results with $(t_{111})_x = 1$. The main axis lengths of the 90% uncertainty ellipsoids are in $[0, 0.018]$ for the camera centers. The rank 0 (smallest), rank $\frac{1}{4}$, rank $\frac{1}{2}$ (median), rank $\frac{3}{4}$ and rank 1 (largest) uncertainty results for points are 0.014, 0.021, 0.029, 0.064, 52, respectively.

Pose Accuracy An other sequence is taken in an indoor controlled environment: the motion of our system is measured on a rail, in a room of dimensions $7m \times 5m \times 3m$. The trajectory is a 1 meter long straight line by translation, with 6 equidistant and aligned poses. The location error (defined in the synthetic experiments) is $E_t = 1.5$ mm.

5. Conclusion

A generic SfM method based on an angular error is introduced, which reduces the drawbacks of previous generic methods. Experiments are done in a non trivial context (a non-central catadioptric and calibrated camera) and include: a ray surface choice discussion, uncertainty and accuracy results, and the automatic/robust/optimal geometry estimation for a long image sequence. Generic (angular) calibration and applications are subjects of future works.

Appendix

We show in this appendix that the composed function $e_{i,j} = arcos(\vec{d}_{i,j}.\vec{D}_{i,j}), \vec{D}_{i,j} = \frac{R_i^\top (\, I_3 \quad -t_i \,) X_j - s_{i,j}}{||R_i^\top (\, I_3 \quad -t_i \,) X_j - s_{i,j}||}$ is not $\mathcal{C}^1$ continuous when $e_{i,j} = 0$. Without loss of generality, we change the space coordinate system such that $\vec{d}_{i,j} = (\, 0 \quad 0 \quad 1 \,)^\top$, and write $\vec{D}_{i,j} = \frac{1}{\sqrt{x^2(\alpha)+y^2(\alpha)+z^2(\alpha)}} (\, x(\alpha) \quad y(\alpha) \quad z(\alpha) \,)^\top$ with $x(\alpha), y(\alpha), z(\alpha)$ three real $\mathcal{C}^1$ continuous functions with parameter α such that $(\, x(0) \quad y(0) \quad z(0) \,) = (\, 0 \quad 0 \quad 1 \,)$. Now, we show that the limit of $\frac{\partial e_{i,j}}{\partial \alpha}$ is not well defined when α converges to 0.

$x(\alpha), y(\alpha), z(\alpha)$ are shortened by x, y, z. The Chain Rule provides $\frac{\partial e_{i,j}}{\partial \alpha} = arcos'(\vec{d}_{i,j}.\vec{D}_{i,j}) \frac{\partial}{\partial \alpha}(\vec{d}_{i,j}.\vec{D}_{i,j})$ with $arcos'(u) = \frac{-1}{\sqrt{1-u^2}}$ and the assumption $|\vec{d}_{i,j}.\vec{D}_{i,j}| < 1$: $\frac{\partial e_{i,j}}{\partial \alpha} = \frac{-1}{x^2+y^2+z^2}(\sqrt{x^2+y^2}\frac{\partial z}{\partial \alpha} - \frac{z}{\sqrt{x^2+y^2}}(x\frac{\partial x}{\partial \alpha} + y\frac{\partial y}{\partial \alpha}))$. Since $(\, x(\alpha) \quad y(\alpha) \quad z(\alpha) \,) \approx (\, \frac{\partial x}{\partial \alpha}(0)\alpha \quad \frac{\partial y}{\partial \alpha}(0)\alpha \quad 1 \,)$, we have $\frac{\partial e_{i,j}}{\partial \alpha} \approx \frac{\alpha}{|\alpha|}\sqrt{(\frac{\partial x}{\partial \alpha})^2(0) + (\frac{\partial y}{\partial \alpha})^2(0)}$. Two $\frac{\partial e_{i,j}}{\partial \alpha}$ limits are obtained: one for each possible α sign.

References

[1] D.G. Aliaga. Accurate Catadioptric Calibration for Realtime pose estimation in Room-size Environments. *ICCV'01*.

[2] "Boujou", *2D3 Ltd*, http://www.2d3.com, 2000.

[3] P. Chang and M. Hebert. Omnidirectional structure from motion. *OMNIVIS'00*.

[4] O.D. Faugeras. Three-Dimensional Computer Vision - A Geometric Viewpoint. *MIT Press*, 1993.

[5] C. Geyer, T. Pajdla, and K. Daniilidis. Courses on Omnidirectional Cameras. *ICCV'03*.

[6] M. Grossberg and S.K. Nayar. A general imaging model and a method for finding its parameters. *ICCV'01*.

[7] R. Haralick, C. Lee, K. Ottenberg and M. Nolle. Review and Analysis of Solutions of the Three Point Perspective Pose Estimation Problem. *IJCV*, 13(3):331-356, 1994.

[8] R. Hartley and A. Zisserman. Multiple View Geometry in Computer Vision. *Cambridge University Press*, 2000.

[9] S.B. Kang. Catadioptric Self-Calibration. *CVPR'00*.

[10] M. Lhuillier. Automatic Structure and Motion using a Catadioptric Camera. *OMNIVIS'05*.

[11] B. Micusik and T. Pajdla. Autocalibration and 3D Reconstruction with Non-Central Catadioptric Cameras. *CVPR'04*.

[12] D. Nister. A Minimal Solution to the Generalized 3-Point Pose Problem. *CVPR'03*.

[13] D. Nister, O. Naroditsky and J. Bergen. Visual Odometry. *CVPR'04*.

[14] R. Pless. Using many cameras as one. *CVPR'03*.

[15] M. Pollefeys, R. Koch and L. Van Gool. Self-Calibration and Metric Reconstruction in spite of Varying and Unknown Internal Camera Parameters. *ICCV'98*.

[16] S. Ramalingam, S.K. Lodha and P. Sturm. A generic Structure-from-Motion Algorithm for Cross Camera Scenarios. *OMNIVIS'04*.

[17] B. Triggs, P.F. McLauchlan, R.I. Hartley and A. Fitzgibbon. Bundle adjustment – a modern synthesis. *Vision Algorithms: Theory and Practice, 2000.*

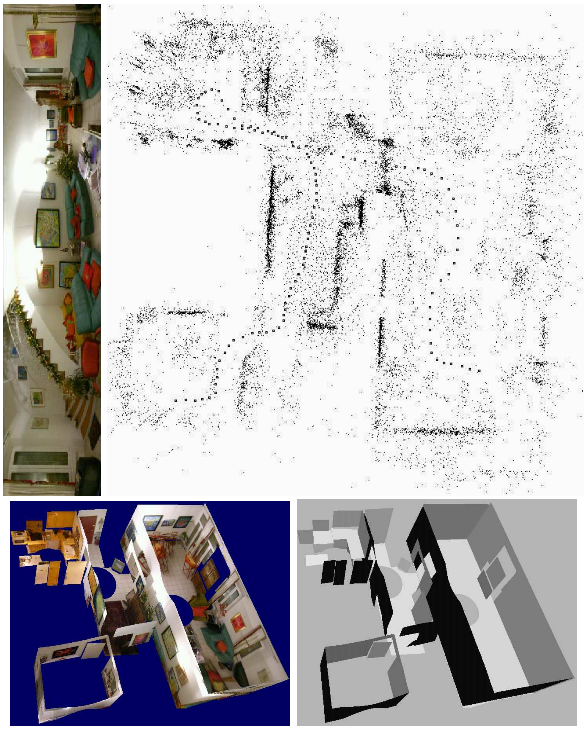

Figure 1. A panoramic image from the House sequence (acquired by a non-central catadioptric camera), a top view of the recovered reconstruction with 112 camera locations (little squares) and 18263 3D points (black points), piecewise planar 3D model from 3D points.