Generic and Real-Time Structure from Motion

E. Mouragnon[1,2], M. Lhuillier[1], M. Dhome[1], F. Dekeyser[2] and P. Sayd[2]

[1]LASMEA UMR 6602, Université Blaise Pascal/CNRS,
63177 Aubière Cedex, France
[2]CEA, LIST, Laboratoire Calculateurs Embarqués et Image,
Boîte Courrier 65, Gif-sur-Yvette, F-91191 France ;

Abstract

We introduce a generic and incremental Structure from Motion method. By generic, we mean that the proposed method is independent of any specific camera model. During the incremental 3D reconstruction, parameters of 3D points and camera poses are refined simultaneously by a generic local bundle adjustment that minimizes an angular error between rays. This method has three main advantages: it is generic, fast and accurate. The proposed method is evaluated by experiments on real data with three kinds of calibrated cameras: stereo rig, perspective and catadioptric cameras.

1 Introduction

The automatic estimation of scene 3D structure and camera motion from an image sequence ("Structure from Motion" or SfM) has been largely studied. Different camera models are used: pinhole, fish-eye, stereo, catadioptric, multi-cameras systems, etc. A lot of **specific** algorithms (i.e. specific to a given camera model) have been successfully developed and are now well known for perspective or stereo rig models [11, 14]. The omni-directional central (catadioptric, fish-eye) or non-central (multi-cameras) systems that offer a larger field of view have also been widely explored [1, 9, 13]. It is a very interesting challenge to develop generic tools for SfM that are exploitable for any camera model. This way has recently been investigated with the introduction of **generic** camera models [7, 17]. In the generic camera model, pixels define image rays in camera coordinate system that can intersect or not in a unique point usually called "projection center". In recent work on generic SfM, camera motion can be estimated using generalization of the classical essential matrix [15, 13] given by Pless Equation [13] and minimal relative pose estimation algorithms [16].

A method is required for the refinement of 3D points and camera poses. The best solution for accuracy is bundle adjustment (BA) [18] applied on all parameters (global BA). However, it is clear that this method can not be real-time. In general, fast SfM methods or Vision-based SLAM [12, 2] (Simultaneous Localization and Mapping) are less accurate than off-line methods where an optimal solution is

BMVC 2007 doi:10.5244/C.21.64

calculated using global BA [18]. In this paper we present a method that makes the most of the accuracy of BA in a generic real-time application. This is possible because we developed an incremental method where not all the 3D structure is refined, but only the lastly estimated parameters at the end of the sequence. In the generic case where different cameras (pinhole, stereo, fish-eye, catadioptric...) are possible, BA is quite different from the classical one used for perspective cameras. Our generic method does not use image projections of a specific camera model but is based on back-projected rays and minimization of an angular error between rays. The first advantage is of course a high ability to change one camera model for another. The second advantage is that the method is effective if the image projection function is not explicit (as in the non-central catadioptric case) and also avoids clustering rays as in [15].

Comparison with previous works on SfM To resume, previous works are:
- generic but not real-time [15].
- real-time but not generic [2, 12], not using bundle adjustment.
- generic thanks to the use of Pless Equation [13, 15] (generalization of the epipolar constraint), but no details are given to solve this equation in common situations.
- using local bundle adjustment but not generic [10, 3], and [3] is not demonstrated in a real-time system with real world data as ours in this paper.

Contributions The first contribution of our work is a generic and real-time SfM method based on an incremental 3D reconstruction and local generic bundle adjustment where an angular error is used. The second contribution is a detailed method to solve Pless Equation (in most cases, it is not a "simple" linear problem as suggested in [13, 15]). We also compare our results with GPS ground truth and with results obtained with the most accurate (but not generic and not real-time) method available: specific global BA.

The remainder of the paper is organized as follows: Section 2 summarizes our approach and the generic camera model. The initialization method and our modified bundle adjustment are respectively explained in Section 3 and 4. Finally, experiments are presented in Section 5.

2 Overview of the Approach

2.1 Camera Model

For any pixel $\mathbf{p}$ of a generic image, the (known) calibration function f of the camera defines an optical ray $r = f(\mathbf{p})$. This projection ray is an oriented line $r = (\mathbf{s}, \mathbf{d})$ where $\mathbf{s}$ is the starting point or origin and $\mathbf{d}$ is the direction of the ray in the camera frame ($||\mathbf{d}|| = 1$). For a central camera, $\mathbf{s}$ is a unique point (camera center) whatever pixel $\mathbf{p}$. In the general case, $\mathbf{s}$ could be any point given by calibration.

2.2 Summary

The method is based on the detection and matching of interest points (Figure 1). In each frame, Harris corners [5] are detected and matched with points detected in a previous frame by computing a Zero Normalized Cross Correlation score in a near region of interest (whatever the kind of camera). The pairs with the highest scores are selected to provide a list of corresponding point pairs between the two images. To ensure a stable estimation of the 3D, a set of frames called key frames are selected. The selection criterion is based on the number of matched points between two consecutive key frames, which must be greater than a constant.

The initialization of the geometry is provided by a method based on the resolution of Pless Equation (Section 3). Then, the algorithm is incremental. For each new video frame, (1) interest points are detected and matched with those of the last key frame (2) the camera pose of the new frame is robustly estimated (Section 3.4) (3) we check if the new frame is selected as a key frame (4) if yes, new 3D points are estimated and a local bundle adjustment (Section 4) is applied.

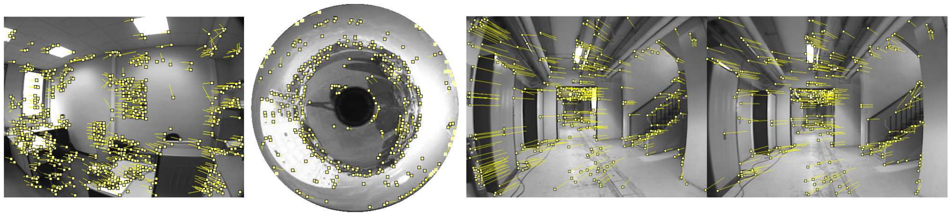

Figure 1: Feature tracks for one image of a generic camera in three cases: perspective (left), catadioptric (middle), and stereo rig (right) cameras.

3 Generic Initialization

3.1 The Pless Equation

Given a set of pixel correspondences between two images, the relative pose $(\mathtt{R}, \mathbf{t})$ of two cameras are estimated in a generic framework. For each 2D points correspondence (x_0, y_0) and (x_1, y_1) between images 0 and 1, we have a correspondence of optical rays $(\mathbf{s}_0, \mathbf{d}_0)$ and $(\mathbf{s}_1, \mathbf{d}_1)$. A ray $(\mathbf{s}, \mathbf{d})$ is defined by its Plücker coordinates $(\mathbf{q}, \mathbf{q}')$ such that $\mathbf{q} = \mathbf{d}$ and $\mathbf{q}' = \mathbf{d} \wedge \mathbf{s}$, which are convenient for this calculation. Let camera 0 be the origin of the global coordinates system and $(\mathtt{R}, \mathbf{t})$ the pose of camera 1 in this frame. The two rays must verify the generalized epipolar constraint (or Pless Equation [13])

$$\mathbf{q}_0'^{\top} \mathtt{R} \mathbf{q}_1 - \mathbf{q}_0^{\top} [\mathbf{t}]_{\times} \mathtt{R} \mathbf{q}_1 + \mathbf{q}_0^{\top} \mathtt{R} \mathbf{q}_1' = 0 \tag{1}$$

where $[\mathbf{t}]_{\times}$ is the skew symmetric cross-product matrix of the 3×1 vector $\mathbf{t}$.

We identify two cases where this equation has an infinite number of solutions. Obviously, this number is infinite if the camera is central (the 3D is recovered up to a scale). We note that Equation 1 is the usual epipolar constraint defined by the essential matrix $\mathtt{E} = [\mathbf{t}]_{\times} \mathtt{R}$ if the camera center is at the origin of the camera frame.

The second case is less obvious but it occurs in practice. In our experiments, we assume that we have only "simple" matches: all projection rays $(\mathbf{s}^i, \mathbf{d}^i)$ of a given

3D point go through a same camera center (in the local coordinate of the generic camera). In other words, we have $\mathbf{q}'_0 = \mathbf{q}_0 \wedge \mathbf{c}^0$ and $\mathbf{q}'_1 = \mathbf{q}_1 \wedge \mathbf{c}^1$ with $\mathbf{c}^0 = \mathbf{c}^1$. For a multi-camera system composed by central cameras (as the stereo rig), it means that 2D points correspondences are only made with points of the same sub-image. This is often the case in practice for two reasons: small regions of interest for reliable matching, or empty intersections between field of views of compositing cameras. If the camera motion is a pure translation ($\mathtt{R} = \mathtt{I}_3$), Equation 1 becomes $\mathbf{q}_0^\top [\mathbf{t}]_\times \mathbf{q}_1 = {\mathbf{q}'_0}^\top \mathbf{q}_1 + \mathbf{q}_0^\top \mathbf{q}'_1 = 0$ where the unknown is $\mathbf{t}$. In this context, the scale of $\mathbf{t}$ can not be estimated. We assume in this work that the camera motion is not a pure translation at the initialization step.

3.2 Solving the Pless Equation

Equation 1 is rewritten as

$$ {\mathbf{q}'_0}^\top \tilde{\mathtt{R}} \mathbf{q}_1 - \mathbf{q}_0^\top \tilde{\mathtt{E}} \mathbf{q}_1 + \mathbf{q}_0^\top \tilde{\mathtt{R}} \mathbf{q}'_1 = 0 \tag{2} $$

where the two 3×3 matrices $(\tilde{\mathtt{R}}, \tilde{\mathtt{E}})$ are the new unknowns. We store the coefficients of $(\tilde{\mathtt{R}}, \tilde{\mathtt{E}})$ in an 18×1 vector $\mathbf{x}$ and see that each value of the 4-tuple $(\mathbf{q}_0, \mathbf{q}'_0, \mathbf{q}_1, \mathbf{q}'_1)$ produces a linear equation $\mathbf{a}^\top \mathbf{x} = 0$. If we have 17 different values of this 4-tuple for each correspondence k, we have 17 equations $\mathbf{a}_k^\top \mathbf{x} = 0$. This is enough to determine $\mathbf{x}$ up to a scale factor [15]. We have built the matrix $\mathtt{A}_{17}$ containing the 17 correspondences such that $\|\mathtt{A}_{17}\mathbf{x}\| = 0$ with $\mathtt{A}_{17}^\top = [\mathbf{a}_1^\top | \mathbf{a}_2^\top | \cdots \mathbf{a}_{17}^\top]$. The resolution depends on the dimension of the $\mathtt{A}_{17}$ kernel which directly depends on the type of camera used. We determine $Ker(\mathtt{A}_{17})$ and its dimension by a Singular Value Decomposition of $\mathtt{A}_{17}$. In this paper, we have distinguished three cases: (1) central cameras with an unique optical center (2) axial cameras with collinear centers and (3) non-axial cameras.

It is not surprising that the kernel dimension of the linear system to solve is greater than one. Indeed, the linear Equation 2 has more unknowns (18 unknowns) than the non-linear Equation 1 (6 unknowns). Possible dimensions are reported in Table 1 and are justified below. Previous works [13, 15] ignored these dimensions, although a (linear) method is heavily dependent on them.

Camera	Central	Axial	Non-Axial
$dim(Ker(\mathtt{A}_{17}))$	10	4	2

Table 1: $dim(Ker(\mathtt{A}_{17}))$ depends on the kind of camera.

Central Camera For central cameras (e.g. pinhole cameras), all optical rays converge at the optical center $\mathbf{c}$. Since $\mathbf{q}'_i = \mathbf{q}_i \wedge \mathbf{c} = [-\mathbf{c}]_\times \mathbf{q}_i$, Equation 2 becomes ${\mathbf{q}_0}^\top ([\mathbf{c}]_\times \tilde{\mathtt{R}} - \tilde{\mathtt{E}} - \tilde{\mathtt{R}}[\mathbf{c}]_\times)\mathbf{q}_1 = 0$. We note that $(\tilde{\mathtt{R}}, \tilde{\mathtt{E}}) = (\tilde{\mathtt{R}}, [\mathbf{c}]_\times \tilde{\mathtt{R}} - \tilde{\mathtt{R}}[\mathbf{c}]_\times)$ is a possible solution of equation 2 for any 3×3 matrix $\tilde{\mathtt{R}}$. Such solutions are "exact": Equation 2 is exactly equal to 0 whatever $(\mathbf{q}_0, \mathbf{q}_1)$. Our "real" solution is $(\tilde{\mathtt{R}}, \tilde{\mathtt{E}}) = (\mathtt{0}, [\mathbf{t}]_\times \mathtt{R})$ if $\mathbf{c} = 0$, and it is not exact due to image noise. Thus the dimension of $Ker(\mathtt{A}_{17})$ is at least $9+1$. Experiments have confirmed that this dimension is 10 (up to noise). In this case, we simply solve the usual epipolar constraint constraint ${\mathbf{q}_0}^\top [\mathbf{t}]_\times \mathtt{R} \mathbf{q}_1 = 0$ as described in [6].

Axial Camera This case includes the common stereo rig of two perspective cameras. Let $\mathbf{c}_a$ and $\mathbf{c}_b$ be two different centers of the camera axis. It is not difficult to prove that "exact" solutions $(\tilde{\mathtt{R}}, \tilde{\mathtt{E}})$ are defined by

$$\tilde{\mathtt{E}} = [\mathbf{c}_a]_\times \tilde{\mathtt{R}} - \tilde{\mathtt{R}}[\mathbf{c}_a]_\times \text{ and } \tilde{\mathtt{R}} \in Vect\{\mathtt{I}_{3\times 3}, [\mathbf{c}_a - \mathbf{c}_b]_\times, (\mathbf{c}_a - \mathbf{c}_b)(\mathbf{c}_a - \mathbf{c}_b)^\top\}$$

based on our assumption of "simple" matches (Section 3.1). Our real solution is not exact due to image noise, and we note that the dimension of $Ker(A_{17})$ is at least 3+1. Experiments have confirmed that this dimension is 4.

We build a basis of 3 exact solutions $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ and a non-exact solution $\mathbf{y}$ with the singular vectors corresponding to the four smallest singular values of $\mathtt{A}_{17}$. The singular values of $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ are 0 (up to computer accuracy) and that of $\mathbf{y}$ is 0 (up to image noise). We calculate the real solution $(\tilde{\mathtt{R}}, \tilde{\mathtt{E}})$ by linear combination of $\mathbf{y}$, $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ such that the resulting matrix $\tilde{\mathtt{R}}$ verifies $\tilde{\mathtt{R}}^\top \tilde{\mathtt{R}} = \lambda \mathtt{I}_{3\times 3}$ or $\tilde{\mathtt{E}}$ is an essential matrix. Let $\mathbf{l}$ be the vector such that $\mathbf{l}^\top = [\lambda_1\ \lambda_2\ \lambda_3]^\top$, and thus we denote as $\tilde{\mathtt{R}}(\mathbf{l})$ and $\tilde{\mathtt{E}}(\mathbf{l})$ the matrix $\tilde{\mathtt{R}}$ and $\tilde{\mathtt{E}}$ extracted from solution $\mathbf{y} - [\mathbf{x}_1|\mathbf{x}_2|\mathbf{x}_3]\mathbf{l}$. Using these notations, we have $\tilde{\mathtt{R}}(\mathbf{l}) = \mathtt{R}_0 - \sum_{i=1}^{3} \lambda_i \mathtt{R}_i$ and $\tilde{\mathtt{E}}(\mathbf{l}) = \mathtt{E}_0 - \sum_{i=1}^{3} \lambda_i \mathtt{E}_i$ with $(\mathtt{R}_i, \mathtt{E}_i)$ extracted from $\mathbf{x}_i$.

Once the basis $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ is calculated, we compute the coordinates of the solution by non-linear minimization of the function $(\lambda, \mathbf{l}) \rightarrow \|\lambda \mathtt{I}_{3\times 3} - \mathtt{R}(\mathbf{l})^\top.\mathtt{R}(\mathbf{l})\|^2$ to obtain $\mathbf{l}$ and thus $\tilde{\mathtt{E}}$. An SVD decomposition is applied to $\tilde{\mathtt{E}}$, and we obtain 4 solutions [6] for $([\mathbf{t}]_\times, \mathtt{R})$. The solution with the minimal epipolar constraint $\|\mathtt{A}_{17}\mathbf{x}\|$ is then selected. Lastly, we refine the 3D scale k by minimizing $k \rightarrow \sum_i ({\mathbf{q}'_{0i}}^\top \mathtt{R}\mathbf{q}_{1i} - \mathbf{q}_{0i}^\top k.[\mathbf{t}]_\times \mathtt{R}\mathbf{q}_{1i} + \mathbf{q}_{0i}^\top \mathtt{R}\mathbf{q}'_{1i})^2$ and perform $\mathbf{t} \leftarrow k\mathbf{t}$.

Non-Axial Camera For a non-axial camera (e.g. a multicamera system with perspective cameras such that centers are not collinear), the problem is also different. In this case, the "exact" solutions are $(\tilde{\mathtt{R}}, \tilde{\mathtt{E}}) \in Vect\{(\mathtt{I}_{3\times 3}, \mathtt{0}_{3\times 3})\}$ based on our assumption of "simple" matches (Section 3.1). The real solution is not exact due to image noise, and we see that the dimension of $Ker(A_{17})$ is at least 1+1. Experiments have confirmed that this dimension is 2. We have not yet experimented this case on real data.

3.3 Initialization with Three Views (RANSAC process)

The first step of the incremental algorithm is the 3D reconstruction of a subsequence containing the first key frames triplet $\{0, 1, 2\}$. A number of random samples are taken, each containing 17 points. For each sample, the relative pose between views 0 and 2 is computed using the abovedescribed method and matched points are triangulated. The pose of camera 1 is estimated with 3D/2D correspondences by iterative refinement minimizing the angular error defined in Section 4.2. The same error is minimized to triangulate points. Finally, the solution producing the highest number of inliers in views 0, 1, and 2 is selected from among all samples. The j-th 3D point is considered as an inlier in view i if the angular error $||\epsilon_j^i||$ is less than ϵ ($\epsilon = 0.01\ rad$ in our experiments).

3.4 Pose Estimates (RANSAC)

The generic pose calculation is useful for both steps of our approach (initialization and incremental process). We assume that the i-th pose $P^i = (\mathtt{R}^i, \mathbf{t}^i)$ of the camera is close to that of the i-1-th pose $P^{i-1} = (\mathtt{R}^{i-1}, \mathbf{t}^{i-1})$. P^i is estimated by iterative non-linear optimization initialized at P^{i-1} with a reduced sample of five 3D/2D correspondences, in conjunction with RANSAC. For each sample, the pose is estimated by minimizing an angular error (Section 4.2) and the overall number of counted inliers (points) includes this pose. The pose with the maximum number of inliers is then selected and another optimization is applied with all inliers.

4 Generic and Incremental Bundle Adjustment

4.1 Definitions

Bundle adjustment (BA) is the refinement of 3D points and camera poses by minimizing a cost function. The number of unknown parameters is 3 for each 3D point and 6 for each camera pose (3 for translation + 3 for rotation). Let $\mathbf{P}_j = [x_j, y_j, z_j, t_j]^\top$ be the homogeneous coordinates of the j-th point in the world frame. Let $\mathtt{R}^i$ and $\mathbf{t}^i$ be the orientation (rotation matrix) and the origin of the i-th camera frame in the world frame.

If $(\mathbf{s}^i_j, \mathbf{d}^i_j)$ is the optical ray corresponding to the observation of $\mathbf{P}_j$ through the i-th camera, the direction of the line defined by $\mathbf{s}^i_j$ and $\mathbf{P}_j$ is $\mathbf{D}^i_j = {\mathtt{R}^i}^\top[\mathtt{I}_3 \mid -\mathbf{t}^i]\mathbf{P}_j - t_j\mathbf{s}^i_j$ in the i-th camera frame. In the ideal case, directions $\mathbf{d}^i_j$ and $\mathbf{D}^i_j$ are parallel (which is equivalent to an image reprojection error of zero pixels).

4.2 Error choice

The classical approach [18, 3] consists in the minimization of a sum of square $||\epsilon^i_j||^2$ where ϵ^i_j is a specific error depending on the camera model (the reprojection error in pixels). In our case, we should minimize a generic error. We define ϵ^i_j as the angle between the directions $\mathbf{d}^i_j$ and $\mathbf{D}^i_j$ defined above.

Some experiments show that convergence of BA is bad with $\epsilon^i_j = \arccos(\mathbf{d}^i_j . \frac{\mathbf{D}^i_j}{||\mathbf{D}^i_j||})$ and satisfactory with ϵ^i_j defined as follows (a theoretical explanation of this is given in [8]). We choose $\epsilon^i_j = \pi(\mathtt{R}^\mathtt{i}_\mathtt{j}\mathbf{D}^i_j)$ with $\mathtt{R}^\mathtt{i}_\mathtt{j}$ a rotation matrix such that $\mathtt{R}^\mathtt{i}_\mathtt{j}\mathbf{d}^i_j = [0\ 0\ 1]^\top$ and π a function $\mathbb{R}^3 \to \mathbb{R}^2$ such that $\pi([x\ y\ z]^\top) = [\frac{x}{z}\ \frac{y}{z}]^\top$. Note that ϵ^i_j is a 2D vector whose Euclidean norm $||\epsilon^i_j||$ is equal to the tangeant of the angle between $\mathbf{d}^i_j$ and $\mathbf{D}^i_j$. The tangeant is a good approximation of the angle if it is small.

4.3 Local Generic Bundle Adjustment

In the incremental 3D reconstruction, when a new key frame I^i is selected, new matched points are triangulated. Then, a stage of optimization is carried out. It is a bundle adjustment or Levenberg-Marquardt minimization of the cost function $f^i(\mathcal{C}^i, \mathcal{P}^i)$ where $\mathcal{C}^i$ and $\mathcal{P}^i$ are respectively the generic camera parameters (extrinsic parameters of key frames) and 3D points chosen for this stage i. As it is

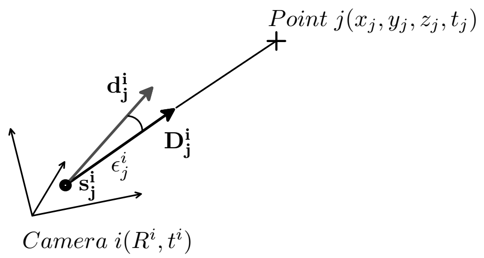

Figure 2: Angular bundle adjustment: the angle between observation ray $(\mathbf{s}_j^i, \mathbf{d}_j^i)$ and 3D ray $\mathbf{D}_j^i$ which goes from $\mathbf{s}_j^i$ to 3D point is minimized.

well known that BA is very time consuming, our idea is to reduce the number of calculated parameters and avoid redundancies in computations. In our modified BA, not all the extrinsic cameras parameters are optimized but only the n last cameras parameters. Coordinates of all 3D points seen in the last n key frames are refined including new points. To bring consistency to the incremental process and ensure that new parameters are compatible with firstly estimated ones, we take account of points reprojections in the N (with $N \geq n$) last frames (typically $n = 3$ and $N = 10$ are good values [10]). Thus, $\mathcal{C}^i$ is the camera list $\{C^{i-n+1} \ \dots \ C^i\}$ and $\mathcal{P}^i$ contains all the 3D points projected on cameras $\mathcal{C}^i$. Cost function f^i is the sum of squared angular errors for all available observations in last key frames $C^{i-N+1} \ \dots \ C^i$ of all 3D points in $\mathcal{P}^i$:

$$f^i(\mathcal{C}^i, \mathcal{P}^i) = \sum_{C^k \in \{C^{i-N+1} \ \dots \ C^i\}} \sum_{p_j \in \mathcal{P}^i} ||\epsilon_j^k||^2.$$

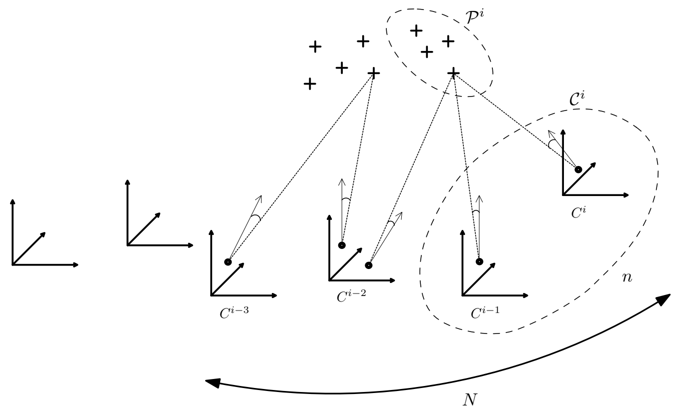

Figure 3: Local angular bundle adjustment when camera C^i is added. Only surrounded points $\mathcal{P}^i$ and cameras $\mathcal{C}^i$ parameters are optimized. Nevertheless, the minimized criterion takes account of 3D points projections in the N last images.

5 Experiments

The incremental generic 3D reconstruction method has been tested on real data with 3 different cameras: a perspective camera, a catadioptric camera and a stereo rig. Examples of frames are available in Figure 1 and sequence characteristics in Table 2. Computation performances are reported on Table 3. In the following experiments, the trajectory obtained with our generic method is compared to GPS ground truth or global specific BA result. A rigid transformation (rotation, translation and scale factor) is applied to the trajectory as described in [4] to fit with reference data. Then, a mean 3D error or 2D error in the horizontal plane can be measured between the generic and the reference trajectory.

5.1 Comparison with Ground Truth (Differential GPS)

The following results are obtained with a pinhole camera embedded on an experimental vehicle equipped with a differential GPS receiver (inch precision). The vehicle trajectory is a “S” of 88 m long (Sequence 1). The calculated motion obtained with our algorithm is compared to data given by the GPS sensor and Figure 4 shows the two trajectories registration. As GPS positions are given in a metric frame we can compare camera locations and measure positioning error in meters: mean 3D error is 1.57 m and 2D error in the horizontal plane is 1.16 m. Computation time is 2 min 32 s for the whole sequence and a mean frame-rate of 6.55 fps.

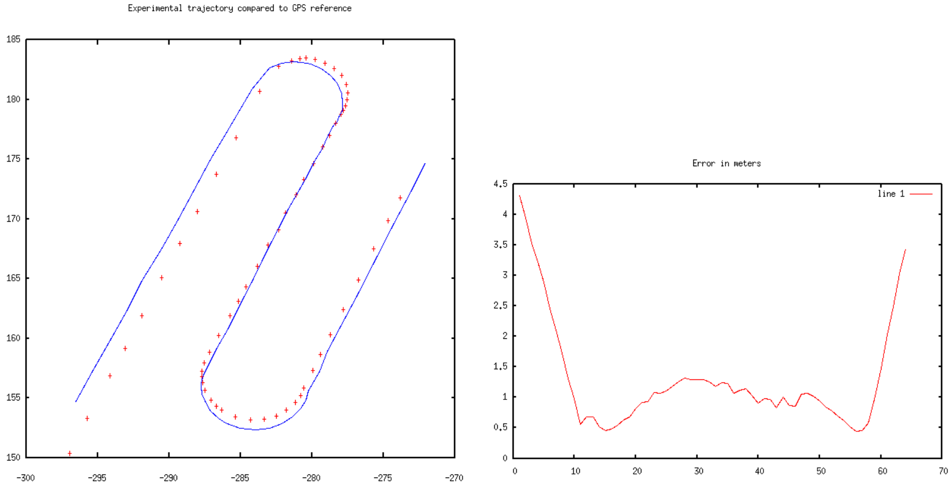

Figure 4: Left: Registration of generic vision trajectory with GPS ground truth. Continuous line represents GPS and points represent vision estimated positions. Right: 3D error (y-axis) along the trajectory (x-axis: key-frame index)

5.2 Comparison with Specific and Global Bundle Adjustment

In the two following examples, ground truth is not available. So, we compare our results with those of the best method available: a global and specific BA (all 3D parameters have been refined so as to obtain an optimal solution with a minimal reprojection error). Sequences characteristics and results are reported on Table 2.

Sequence 2 is taken in an indoor environment with a hand-held pinhole camera. A very accurate result is obtained: the mean 3D error is less than 6.5 *cm* for a trajectory length of (about) 15 *m*. The relative error is 0.45%.

Sequence 3 is taken in an outdoor environment with a hand-held catadioptric camera (the 0-360 mirror with the Sony HDR-HC1E camera visible on Figure 5, DV format). The useful part of the rectified image is contained in a circle whose diameter is 458 pixels. The accuracy is also good: the mean 3D error is less than 9.2 *cm* for a trajectory length of (about) 40 *m*. The relative error is 0.23%.

Sequence 4 is taken with a stereo rig (baseline: 40 *cm*) in a corridor (Figure 5). The image is composed of two sub-images of 640 × 480 *pix*. The trajectory (20 *m* long) is compared to results obtained with left/right camera and global BA. The mean 3D error is 2.7/8.4 *cm* compared to left/right camera and the relative error is 0.13/0.42%.

Sequence	Camera	#Frames	#Key frames	#3D Pts	#2D Pts	Traj. length
Sequence 1	pinhole	996	66	4808	17038	88 m
Sequence 2	pinhole	511	48	3162	11966	15 m
Sequence 3	catadioptric	1493	132	4752	18198	40 m
Sequence 4	stereo rig	303	28	3642	14189	20 m

Table 2: Characteristics of video sequences.

Camera	Image size	Detection+Matching	Frame	Key frame	Mean rate
Pinhole	512 × 384	0.10	0.14	0.37	**6.3 fps**
Catadioptric	464 × 464	0.12	0.15	0.37	**5.9 fps**
Stereo rig	1280 × 480	0.18	0.25	0.91	**3.3 fps**

Table 3: Computation times in *seconds* for our three cameras (detection and matching are included in Frame or Key frame times)

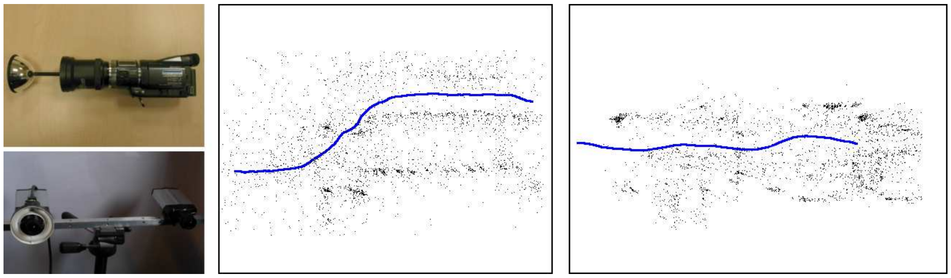

Figure 5: Left: catadioptric camera and stereo rig. Middle and right: top views of 3D reconstructions for Sequence 3 (middle) and Sequence 4 (right). Trajectory in blue and 3D points in black.

6 Conclusion

We have developped and experimented a generic method for the real-time Structure from Motion problem. We presented a complete process that starts with a generic initialization followed by an incremental 3D reconstruction of the scene and camera motion. The accuracy is brought by a local bundle adjustment minimizing an angular error. Experiments proved that it is easy to change one camera model for another, and promising results have been obtained on real data with three different kinds of cameras. Now, we are interested in experimenting our approach on more complex multi-camera systems.

References

[1] P. Chang and M. Hebert. Omni-directional structure from motion. In *Proc. of the IEEE Workshop on Omnidirectional Vision*, 2000.

[2] A. Davison. Real-time simultaneous localization and mapping with a single camera. In *Proc. of ICCV*, 2003.

[3] C. Engels, H. Stewénius, and D. Nistér. Bundle adjustment rules. In *Photogrammetric Computer Vision*, September 2006.

[4] O.D. Faugeras and M. Hebert. The representation, recognition, and locating of 3-D objects. *IJRR*, 5(3): pp 27-52, 1986.

[5] C. Harris and M. Stephens. A combined corner and edge detector. In *4th ALVEY Vision Conference*, 1988.

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

[7] J. Kannala and S.S. Brandt. A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *PAMI*, 28(8): pp 1335-1340, 2006.

[8] M. Lhuillier. Automatic Scene Structure and Camera Motion using a Catadioptric System. *CVIU 2007*, doi: 10.1016/j.cviu.2007.05.004 (to appear).

[9] B. Micusik and T. Pajdla. Autocalibration & 3D reconstruction with non-central catadioptric cameras. In *Proc. of CVPR*, 2004.

[10] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd. Real time localization and 3D reconstruction. In *Proc. of CVPR*, June 2006.

[11] D. Nister. An efficient solution to the five-point relative pose problem. *PAMI*, 26(6): pp 756-777, 2004.

[12] D. Nister, O. Naroditsky, and J Bergen. Visual odometry. In *Proc. of CVPR*, 2004.

[13] R. Pless. Using many cameras as one. In *Proc. of CVPR*, 2003.

[14] M. Pollefeys, R. Koch, M. Vergauwen, and L. Van Gool. Automated reconstruction of 3D scenes from sequences of images. *Isprs Journal Of Photogrammetry And Remote Sensing*, 55(4): pp 251-267, 2000.

[15] S. Ramalingam, S. Lodha, and P. Sturm. A generic structure-from-motion framework. *CVIU*, 103(3): pp 218-228, 2006.

[16] H. Stewénius, D. Nistér, M. Oskarsson, and K. Åström. Solutions to minimal generalized relative pose problems. In *Workshop on Omnidirectional Vision*, 2005.

[17] P. Sturm and S. Ramalingam. A generic concept for camera calibration. In *Proc. of ECCV*, 2004.

[18] Bill Triggs, Philip McLauchlan, Richard Hartley, and Andrew Fitzgibbon. Bundle adjustment – A modern synthesis. In *Vision Algorithms: Theory and Practice.* 2000.