Motion Estimation for Hybrid Cameras using Point and Line Feature Fusion

Sang Ly [1], Cédric Demonceaux [2], Pascal Vasseur [3] and Claude Pégard [1]
[1] MIS Laboratory. University of Picardie Jules-Verne. Amiens. France
[2] Le2i - UMR CNRS 5158. University of Burgundy. Le Creusot. France
[3] LITIS Laboratory. University of Rouen. Rouen. France

Abstract

We present in this paper a structure-from-motion approach for single viewpoint cameras using point and line features. From the image sequence captured by any central projection cameras, the motion can be recovered by decoupling of orientation and displacement: rotation is estimated from vanishing points of parallel lines and translation is calculated from known rotation and point/line correspondences. Next, the 3D structure is reconstructed by triangulation of projective rays or planes. Lastly, the camera motion and 3D scene are optionally refined by bundle adjustment. This approach can be applied to recover the motion of autonomous robots or the arrangement of a vision-based surveillance system equipped with any single viewpoint cameras such as perspective, fish-eye and catadioptric cameras. Moreover, the translation can be estimation from points and/or lines depending on the availability of these features in the images. The proposed algorithm has been validated on simulation data and real images.

1 Introduction

Vision-based system has recently become a widely used assisting device in the navigation of autonomous robots besides the conventional ones such as Global Positioning System (GPS) and Inertial Navigation System (INS). Recently, researchers have utilized many camera types in order to exploit their advantageous characteristics such as the wide field of view of omnidirectional cameras and the uniform spatial resolution of perspective cameras. To make use of such a heterogeneous sensor mixture, we propose in this paper a structure-from-motion algorithm using image correspondences across multiple hybrid views. A question arisen here is that what kind of image features should be used. Between point and line features, the correspondence task is more trivial for lines than for points over multiple views of heterogeneous cameras. Lines is less likely to be produces by noise than points in man-made environment. Lines are less numerous but more informative. Lines are less affected by occlusions as each line can be reconstructed from its different segments in multiple images. However, the disadvantage of using lines is that the reconstruction is only feasible from at least three views whereas it can be done using point correspondences across two views. Moreover, in some particular scenes, one feature is more dominant than the other. Therefore, we propose a structure-from-motion method using any available point or line features. The following sections introduce some related works on structure-from-motion problem.

1.1 Structure from motion using point features

Point-based structure-from-motion methods may be started with *factorization* technique [12, 19] in which the camera motion and scene structure are "factorized" from the image feature matrix.

Besides factorization, the *minimal structure from motion solutions* such as 8-point [6] or 5-point [14] algorithms have been proposed to recover the camera pose from image correspondences.

Recently, *L∞ optimization* methods have been developed to solve the structure-from-motion problem. This approach is based on second-order cone programming (SOCP) to estimate the camera translations and 3D points assuming known rotations [9, 13, 16].

1.2 Structure from motion using line features

There exist numerous works on structure from motion using straight lines [2, 4, 7, 17, 18, 22].

Line-based structure-from-motion algorithms can be classified firstly to *factorization* technique. Without any assumption about camera calibration and 3D information, the camera and line locations are recovered up to projective [11, 20] or affine [15] transformation.

Secondly, there exist *sequential approaches* which are composed of three stages: (i) camera motion estimation, (ii) feature triangulation to obtain 3D structure and (iii) optimization by bundle adjustment [8, §18]. In the first stage, camera transformations can be recovered using matching tensor built up from line correspondences in triplets of views [5,7,21]. Concerning the second and last stages, the principle difference among the previous works is in the parameterization of 3D lines. A thorough description of line representations and their characteristics can be found in [2].

We propose in this paper a sequential structure-from-motion approach assuming calibrated cameras. Firstly, the camera transformations are recovered by decoupling of rotations and translations: rotations are estimated from vanishing points of parallel lines, and translations are linearly calculated from point/line correspondences. Secondly, the 3D structure is reconstructed by the triangulation of projection rays and planes. Finally and also optionally, the camera motion and scene structure are optimized by bundle adjustment. Throughout three stages of the proposed algorithm, we represent our calibrated single viewpoint cameras by the generic spherical camera model to permit an application to several kinds of cameras such as perspective, central catadioptric and fish-eye cameras [1, 23]. Moreover, we introduce a linear translation estimation using point and/or line correspondences, which permits a flexible use depending on the visibility of these features across multiple views. The next three sections describe our estimation algorithm in three stages. We show then the experimental results obtained with simulation and real images; and lastly some conclusions.

2 Linear motion estimation

In this section, we present a linear motion estimation approach by decoupling of rotation and translation: rotations are estimated from vanishing points of parallel lines and translations are recovered from known rotations, point and line features.

Notation: Matrices are denoted with Sans Serif fonts, vectors with bold fonts and scalars with italics.

Consider three central cameras $\mathbf{C}_i$ $(i = 1...3)$ as illustrated in figure 1. Let the coordinate system origin be at the first camera center and $[\mathsf{R}_i|\mathbf{t}_i]$ represent the [Rotation|translation] between $\mathbf{C}_i$ and the origin, hence $[\mathsf{R}_1|\mathbf{t}_1] = [\mathrm{I}|\mathbf{0}]$.

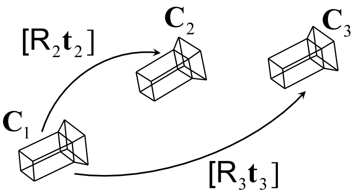

Figure 1. Camera transformations

2.1 Rotation estimation

Rotation between two cameras can be estimated from correspondences of vanishing points, i.e. the intersection of parallel lines [3]. After the detection of vanishing points $\mathbf{V}_i$ in all images, each rotation R_i can be recovered linearly as follows:

$$\mathbf{V}_i = \mathsf{R}_i \mathbf{V}_1 \tag{1}$$

2.2 Translation estimation

We demonstrate in this section that translations can be recovered using point correspondences between two views and line correspondences among three views.

Bilinear constraint of point correspondences: Point correspondences in two views (a, b) related by the transformation $[\mathsf{R}|\mathbf{t}]$ satisfy the following constraint [3]:

$$(\mathsf{R}\mathbf{p}_a \times \mathbf{p}_b)^T \mathbf{t} = 0 \tag{2}$$

Trilinear constraint of line correspondences: Each line in image plan can be back-projected on the sphere as a great circle associated with a normal. A normal correspondence across three views $(1, a, b)$ can be related through the transformations among these views [10]:

$$[\mathbf{n}_1]_\times \mathsf{R}_a^T \mathbf{n}_a \mathbf{n}_b^T \mathbf{t}_b - [\mathbf{n}_1]_\times \mathsf{R}_b^T \mathbf{n}_b \mathbf{n}_a^T \mathbf{t}_a = 0 \tag{3}$$

Given three cameras $\mathbf{C}_i$, the constraints (2) and (3) can be encapsulated in the following linear system which permits the translation estimation from any available point or line correspondences:

$$\mathsf{A}\mathbf{X} = 0 \tag{4}$$

where

$$\mathsf{A} = \begin{bmatrix} (\mathsf{R}_a \mathbf{p}_1 \times \mathbf{p}_a)^T & 0 \\ 0 & (\mathsf{R}_b \mathbf{p}_1 \times \mathbf{p}_b)^T \\ -(\mathsf{R}_b \mathsf{R}_a^T \mathbf{p}_a \times \mathbf{p}_b)^T \mathsf{R}_b \mathsf{R}_a^T & (\mathsf{R}_b \mathsf{R}_a^T \mathbf{p}_a \times \mathbf{p}_b)^T \\ -[\mathbf{n}_1]_\times \mathsf{R}_b^T \mathbf{n}_b \mathbf{n}_a^T & [\mathbf{n}_1]_\times \mathsf{R}_a^T \mathbf{n}_a \mathbf{n}_b^T \end{bmatrix}$$

and

$$\mathbf{X} = (\mathbf{t}_a^T \mathbf{t}_b^T)^T$$

Note that the third row of matrix A is linearly dependent of the first and second rows. However, in case of noisy data, we can still use this relation for the estimation without redundancy.

2.3 Reconstruction

Each 3D point $\mathbf{P}$ is reconstructed by the triangulation of the projection rays passing through $\mathbf{P}$, $\mathbf{C}_i$ and the image point projected on the sphere $\mathbf{p}_i$ (5). The linear solution of $\mathbf{P}$ is given in (6).

$$\mathbf{P} = \mathbf{C}_i + \alpha_i(\mathbf{p}_i - \mathbf{C}_i) \tag{5}$$

$$\mathsf{B}\hat{\mathbf{P}} = \mathbf{C} \tag{6}$$

where

$$\mathsf{B} = \begin{bmatrix} \mathrm{I} & \mathbf{C}_1 - \mathbf{p}_1 & 0 & 0 \\ \mathrm{I} & 0 & \mathbf{C}_2 - \mathbf{p}_2 & 0 \\ \mathrm{I} & 0 & 0 & \mathbf{C}_3 - \mathbf{p}_3 \end{bmatrix},$$
$$\hat{\mathbf{P}} = (\mathbf{P}^T, \alpha_1, \alpha_2, \alpha_3)^T \text{ and } \mathbf{C} = (\mathbf{C}_1^T, \mathbf{C}_2^T, \mathbf{C}_3^T)^T$$

Each 3D line is reconstructed by the intersection of the projective planes passing through line correspondences across three views:

$$\mathsf{G}\hat{\mathbf{L}} = 0 \tag{7}$$

where

$$\mathsf{G} = \begin{bmatrix} \mathbf{n}_1^T & 0 \\ \mathbf{n}_2^T \mathsf{R}_2 & \mathbf{n}_2^T \mathbf{t}_2 \\ \mathbf{n}_3^T \mathsf{R}_3 & \mathbf{n}_3^T \mathbf{t}_3 \end{bmatrix} \text{ and } \hat{\mathbf{L}} = (\mathbf{L}^T, 1)^T$$

From the singular value decomposition of $\mathsf{G} = \mathsf{UDV}^T$, the two columns of V corresponding to two largest singular values can be used to define the line intersection of the planes [8, §12.7].

2.4 Bundle adjustment

This optional optimization stage refines the camera motion and 3D structure by minimizing the reprojection error of points and lines on spherical images.

Each camera is parameterized by the 7-vector $\mathbf{c}_i = (r_0, r_1, r_2, r_3, t_x, t_y, t_z)_i$ where (r_0, r_1, r_2, r_3) is the quaternion representation of the rotation and (t_x, t_y, t_z) the conventional translation. Each 3D point is described by the 3-vector $\mathbf{p}_j = (p_x, p_y, p_z)_j$. Each 3D line is represented by the 6-vector $\mathbf{l}_k = (e_x^1, e_y^1, e_z^1, e_x^2, e_y^2, e_z^2)_k$ established by two points e^1 and e^2 on the line.

The parameter vector in the optimization is defined by all parameters describing i cameras, j points and k lines $\mathbf{Q} = (\mathbf{c}_1 \dots \mathbf{c}_i, \mathbf{p}_1 \dots \mathbf{p}_j, \mathbf{l}_1 \dots \mathbf{l}_k)$.

Bundle adjustment minimizes the following reprojection error with respected to all camera, 3D line and point parameters:

$$\sum_i \sum_j d_p(\hat{\mathbf{p}}_{ij}, \mathtt{P}(\mathbf{c}_i, \mathbf{p}_j)) + \sum_i \sum_k d_l(\hat{\mathbf{n}}_{ik}, \mathtt{P}(\mathbf{c}_i, \mathbf{l}_k)) \quad (8)$$

where

$$d_p = \hat{\mathbf{p}}_{ij} \times (\mathsf{R}_i \mathbf{p}_j + \mathbf{t}_i)$$
$$d_l = \hat{\mathbf{n}}_{ik} \times [(\mathsf{R}_i \mathbf{e}_k^1 + \mathbf{t}_i) \times (\mathsf{R}_i \mathbf{e}_k^2 + \mathbf{t}_i)]$$

where $\mathtt{P}$ is the spherical projection of point $\mathbf{p}_j$ or line $\mathbf{l}_k$ in camera $\mathbf{c}_i$. $\hat{\mathbf{p}}_{ij}$ and $\hat{\mathbf{n}}_{ik}$ are the spherical back-projection of the image point j and the image line k in camera i respectively.

The minimization can be solved by Levenberg-Marquardt non-linear algorithm. The initial parameter estimate $\mathbf{Q}_0$ is provided by the camera motion recovery and reconstruction stages.

Each row of the Jacobian matrix is calculated for each point and line in each camera:
Points:

$$\frac{\partial d_p}{\partial \mathbf{Q}} = [\frac{\partial d_p}{\partial \mathbf{c}_1} \dots \frac{\partial d_p}{\partial \mathbf{c}_i}, \frac{\partial d_p}{\partial \mathbf{p}_1} \dots \frac{\partial d_p}{\partial \mathbf{p}_j}, 0 \dots 0] \quad (9)$$

Lines:

$$\frac{\partial d_l}{\partial \mathbf{Q}} = [\frac{\partial d_l}{\partial \mathbf{c}_1} \dots \frac{\partial d_l}{\partial \mathbf{c}_i}, 0 \dots 0, \frac{\partial d_l}{\partial \mathbf{l}_1} \dots \frac{\partial d_l}{\partial \mathbf{l}_k}] \quad (10)$$

3 Experimental results

3.1 Simulation

We first create a set of 10 points and 10 lines randomly distributed in a sphere with 5 meter radius. Three cameras with an average baseline of 0.5 meter observe these features at a distance of 10 meters. The translations among 3 cameras are recovered from points using 5-point algorithm [14] and from both points and lines using our approach.

Points and line normals are on unitary spheres, thus may be specified by elevation and azimuth angles. Gaussian noise of zero mean and varying standard deviations is added to each angle of every point and normal. Figure 2 shows the average angular error of all translations after 1000 runs. It can be seen that our linear estimation (in green) is more robust to noise than 5-point estimation (in red) and moreover the bundle adjustment stage (in blue) optimizes the linear solution.

3.2 The door sequence

In this section, we evaluate different motion estimation approaches based on: i. points (5-point algorithm [14]), ii. lines (our approach) and iii. combination of points and lines (our approach).

Three image samples captured by two fish-eye ($\mathbf{C}_1$ and $\mathbf{C}_2$) and one perspective ($\mathbf{C}_3$) cameras are illustrated in figure 3. From 15 point and 13 line correspondences across these images, the camera motion and 3D structure are recovered and refined by bundle adjustment which converges after 5 iterations. A snapshot of the reconstruction is shown in figure 5.

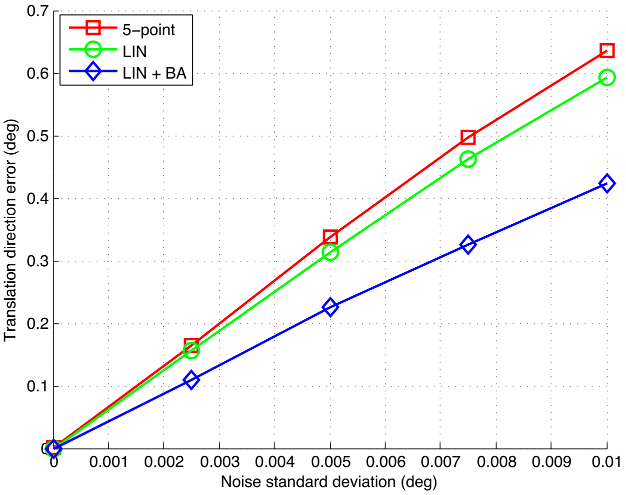

Figure 2. Estimation error of our approach compared to 5-point algorithm

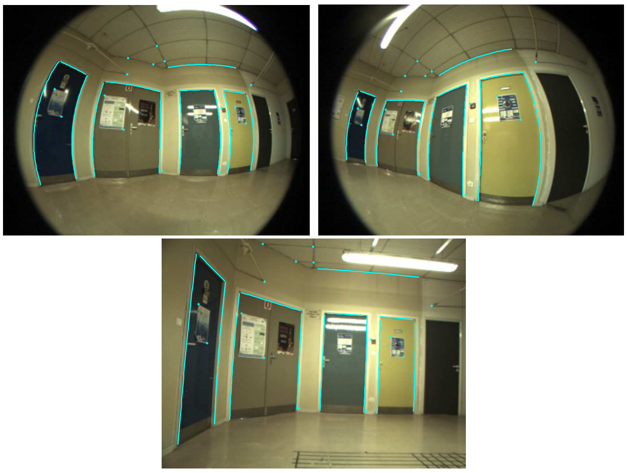

Figure 3. Fish-eye and perspective images with point and line correspondences

To evaluate the up-to-scale structure reconstruction, we verify the dimension of the reconstructed doors (table 1). Four doors (with extracted borders in figure 3) from left to right are denoted Door 1 to 4 respectively. Using height/width ratio of each door obtained from the reconstruction, we deduce its height given its real width. The result is not satisfied for the first door as it is near the image border where there is much distortion, especially in fish-eye images. The line-based approaches provide much better results than point-based-only method and there is no important difference between the estimation using only lines and the estimation by combining points and lines. The reason of this may be that line-based estimation suffers the effect of noise less than point-based one, and consequently adding point feature does not improve significantly the result of line-based estimation.

The re-projection of 3D lines into one of the fish-eye views is illustrated in figure 4. As can be seen from this figure, the point-based approach is very sensitive to noise whereas line-based and point-line-combining approaches perform well in the presence of noise and do not differ from each other.

	Door 1	Door 2	Door 3	Door 4
Points	242	199	215	212
Lines	228	206	206	200
Points+Lines	229	206	206	204
Real height	203	203	203	203

Table 1. Structure reconstruction result

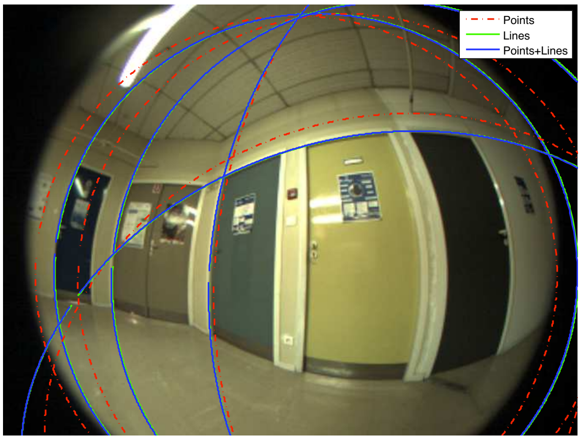

Figure 4. Reprojection of reconstructed 3D lines

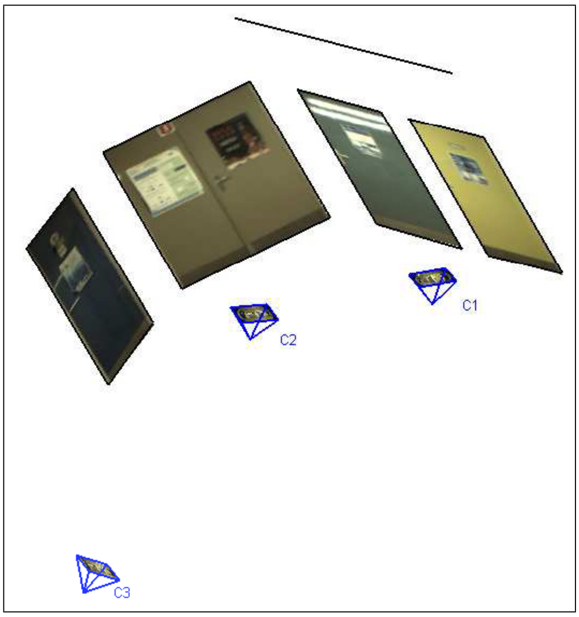

Figure 5. Reconstruction

4 Conclusions

We have proposed in this paper a structure-from-motion algorithm for single viewpoint cameras using point and line correspondences. The presented technique has been validated on simulation data and with real images. To recover the transformations among multiple views, we estimate the rotations using vanishing points and the translations from known rotations and point/line correspondences by a linear algorithm. The proposed algorithm can be applied to any type of single viewpoint cameras and moreover, the translations can be recovered from any available point and/or line features.

References

[1] J.P. Barreto and H. Araujo, *Issues on the geometry of central catadioptric image formation*, CVPR, 2001, pp. II:422–427.

[2] A.E. Bartoli and P.F. Sturm, *Structure-from-motion using lines: Representation, triangulation, and bundle adjustment*, CVIU **100** (2005), no. 3, 416–441.

[3] J.C. Bazin, C. Demonceaux, P. Vasseur, and I. Kweon, *Motion estimation by decoupling rotation and translation in catadioptric vision*, CVIU **114** (2010), no. 2, 254–273.

[4] M. Chandraker, J.W. Lim, and D. Kriegman, *Moving in stereo: Efficient structure and motion using lines*, ICCV, 2009, pp. 1741–1748.

[5] A.W. Fitzgibbon and A. Zisserman, *Automatic camera recovery for closed or open image sequences*, ECCV, 1998, p. I: 311.

[6] R.I. Hartley, *In defense of the eight-point algorithm*, PAMI **19** (1997), no. 6, 580–593.

[7] ———, *Lines and points in three views and the trifocal tensor*, IJCV **22** (1997), no. 2, 125–140.

[8] R.I. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, Cambridge University Press, June 2004.

[9] F. Kahl and R.I. Hartley, *Multiple-view geometry under the l∞-norm*, PAMI **30** (2008), no. 9, 1603–1617.

[10] S. Ly, C. Demonceaux, and P. Vasseur, *Translation estimation for single viewpoint cameras using lines*, ICRA, 2010, pp. 1928–1933.

[11] D. Martinec and T. Pajdla, *Line reconstruction from many perspective images by factorization*, CVPR, 2003, pp. I: 497–502.

[12] ———, *3d reconstruction by fitting low-rank matrices with missing data*, CVPR05, 2005, pp. I: 198–205.

[13] ———, *Robust rotation and translation estimation in multiview reconstruction*, CVPR07, 2007, pp. 1–8.

[14] D. Nistér, *An efficient solution to the five-point relative pose problem*, IEEE Trans. Pattern Anal. Mach. Intell. **26** (2004), 756–777.

[15] L. Quan and T. Kanade, *Affine structure from line correspondences with uncalibrated affine cameras*, PAMI **19** (1997), no. 8, 834–845.

[16] K. Sim and R. Hartley, *Recovering camera motion using l-inf minimization*, CVPR06, 2006, pp. I: 1230–1237.

[17] M.E. Spetsakis and Y. Aloimonos, *Structure from motion using line correspondences*, IJCV **4** (1990), no. 3, 171–183.

[18] C.J. Taylor and D.J. Kriegman, *Structure and motion from line segments in multiple images*, PAMI **17** (1995), no. 11, 1021–1032.

[19] C. Tomasi and T. Kanade, *Shape and motion from image streams under orthography: A factorization method*, IJCV **9** (1992), no. 2, 137–154.

[20] B. Triggs, *Factorization methods for projective structure and motion*, CVPR, 1996, pp. 845–851.

[21] ———, *Linear projective reconstruction from matching tensors*, IVC **15** (1997), no. 8, 617–625.

[22] J.Y. Weng, T.S. Huang, and N. Ahuja, *Motion and structure from line correspondences: Closed-form solution, uniqueness, and optimization*, PAMI **14** (1992), no. 3, 318–336.

[23] X.H. Ying and Z.Y. Hu, *Can we consider central catadioptric cameras and fisheye cameras within a unified imaging model*, ECCV, 2004, pp. Vol I: 442–455.