Linear Global Translation Estimation with Feature Tracks

Zhaopeng Cui[1]
zhpcui@gmail.com
Nianjuan Jiang[2]
nianjuan.jiang@adsc.com.sg
Chengzhou Tang[1]
chengzhout@gmail.com
Ping Tan[1]
pingtan@sfu.ca

[1] GrUVi Lab
Simon Fraser University
Burnaby, Canada

[2] Advanced Digital Sciences Center of Illinois
Singapore

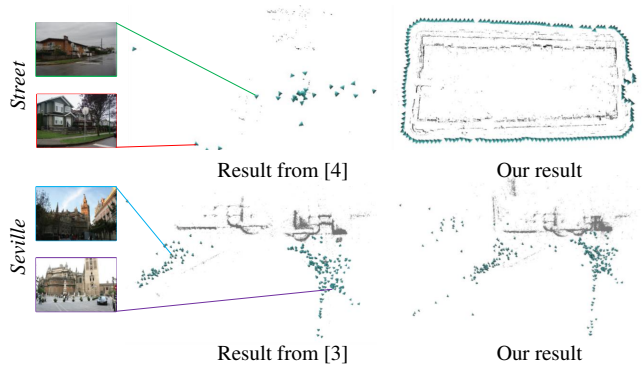

Figure 1: 1DSFM [4] and triplet-based methods (*e.g.* [3]) require strong association among images. As shown in the left, they fail for images with weak association. In comparison, as shown in the right, the results from our method do not suffer from such problems.

Global structure-from-motion (SfM) algorithms register all cameras simultaneously, which are potentially more efficient and less prone to drifting than incremental SfM methods. Global SfM methods often solve the camera orientations and positions separately. This paper focuses on the problem of global position (*i.e.* translation) estimation.

Essential matrix based global translation estimation methods (*e.g.* [1]) usually degenerate at collinear camera motion because the translation scale is not determined by an essential matrix. Trifocal tensor based methods (*e.g.* [3]) usually rely on a strongly connected camera-triplet graph, where two triplets are connected by their common edge. The 3D reconstruction will distort or break into disconnected components when such strong association among images does not exist. The recent 1DSfM method [4] designs a smart filter to discard outlier essential matrices and solves scene points and cameras together by enforcing orientation consistency. However, this method requires abundant association between input images, *e.g.* $\sim O(n^2)$ essential matrices for n cameras, which is more suitable for Internet images and often fails on sequentially captured data.

The data association problem of [4] and [3] is exemplified in Figure 1. The *Street* example on the top is a sequential data where each image is only matched upto 4 neighbors. 1DSfM fails on this example due to insufficient image association. In the *Seville* example on the bottom, those Internet images are mostly captured from two viewpoints (see the two representative sample images) with weak affinity between images at different viewpoints. This weak data association causes seriously distorted reconstruction for the triplet-based method in [3].

This paper introduces a direct linear algorithm to address the presented challenges. It avoids degeneracy at collinear motion and deals with weakly associated data. Our method capitalizes on constraints from essential matrices and feature tracks. As shown in Figure 2 (a), the location of a scene point $\mathbf{p}$ can be computed as the middle point of the mutual perpendicular line segment AB of the two rays passing through $\mathbf{p}$'s image projections:

$$\mathbf{p} = \frac{1}{2}(A+B) = \frac{1}{2}(\mathbf{c}_i + s_i\mathbf{m}_i + \mathbf{c}_j + s_j\mathbf{m}_j). \tag{1}$$

Here, $\mathbf{c}_i$ and $\mathbf{c}_j$ are the two camera centers. The two unit vectors $\mathbf{m}_i$ and $\mathbf{m}_j$ origin from the camera centers and point toward the image projections of $\mathbf{p}$. s_i and s_j are the distances from the points A, B to $\mathbf{c}_i, \mathbf{c}_j$ respectively.

We use the rotation trick in [3] to compute $\mathbf{m}_i$ and $\mathbf{m}_j$ by rotating the relative translation direction $\mathbf{c}_{ij}$ between $\mathbf{c}_i$ and $\mathbf{c}_j$, *i.e.* $\mathbf{m}_i = \mathbf{R}(\theta_i)\mathbf{c}_{ij}$ and $\mathbf{m}_j = -\mathbf{R}(\theta_j)\mathbf{c}_{ij}$. Then Equation 1 becomes

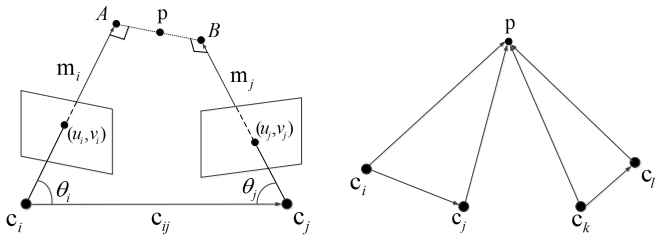

Figure 2: (a) The positions of a scene point $\mathbf{p}$ and two camera centers $\mathbf{c}_i$ and $\mathbf{c}_j$ satisfy a linear constraint. (b) The positions of four cameras seeing the same scene point satisfy a linear constraint.

$$\mathbf{p} = \frac{1}{2}\left(\mathbf{c}_i + s_i\mathbf{R}(\theta_i)\frac{\mathbf{c}_j - \mathbf{c}_i}{\|\mathbf{c}_j - \mathbf{c}_i\|} + \mathbf{c}_j + s_j\mathbf{R}(\theta_j)\frac{\mathbf{c}_i - \mathbf{c}_j}{\|\mathbf{c}_i - \mathbf{c}_j\|}\right). \tag{2}$$

The two 3D rotation matrices $\mathbf{R}(\theta_i)$ and $\mathbf{R}(\theta_j)$ rotate the relative translation direction $\mathbf{c}_{ij}$ to the directions $\mathbf{m}_i$ and $\mathbf{m}_j$. Both rotations can be computed easily in the local pairwise reconstruction. In addition, the two ratios $s_i/\|\mathbf{c}_j - \mathbf{c}_i\|$ and $s_j/\|\mathbf{c}_j - \mathbf{c}_i\|$ can be computed by the middle-point algorithm [2]. Thus, Equation 2 is reduced to,

$$\mathbf{p} = \frac{1}{2}\left((\mathbf{A}_j^{ij} - \mathbf{A}_i^{ij})(\mathbf{c}_i - \mathbf{c}_j) + \mathbf{c}_i + \mathbf{c}_j\right) \tag{3}$$

where $\mathbf{A}_i^{ij} = s_i/||\mathbf{c}_j - \mathbf{c}_i||\mathbf{R}(\theta_i)$ and $\mathbf{A}_j^{ij} = s_j/||\mathbf{c}_i - \mathbf{c}_j||\mathbf{R}(\theta_j)$ are known matrices. This equation provides a linear constraint among positions of two camera centers and a scene point.

If the same scene point $\mathbf{p}$ is visible in two image pairs $\mathbf{c}_i, \mathbf{c}_j$ and $\mathbf{c}_k, \mathbf{c}_l$ as shown in Figure 2 (b), we obtain two linear equations about $\mathbf{p}$'s position according to Equation 3. We can eliminate $\mathbf{p}$ from these equations to obtain a linear constraint among four camera centers as the following,

$$(\mathbf{A}_j^{ij} - \mathbf{A}_i^{ij})(\mathbf{c}_i - \mathbf{c}_j) + \mathbf{c}_i + \mathbf{c}_j = (\mathbf{A}_l^{kl} - \mathbf{A}_k^{kl})(\mathbf{c}_k - \mathbf{c}_l) + \mathbf{c}_k + \mathbf{c}_l. \tag{4}$$

Given a set of images, we build feature tracks and collect such linear equations from camera pairs on the same feature track. Solving these equations will provide a linear global solution of camera positions. To resolve the gauge ambiguity, we set the orthocenter of all cameras at origin when solving these equations.

This direct linear method minimizes a geometric error, which is the Euclidean distance between the scene point to its corresponding rays of projection. A key finding in this paper is that, a direct linear solution (without involving scene points) exists by minimizing the point-to-ray error instead of the reprojection error. Since the point-to-ray error approximates the reprojection error well when cameras are calibrated, our method is a good linear initialization for the final nonlinear BA. At the same time, we minimize the L_1 norm when solving the linear equation of camera positions. We derive a linearization of the alternating direction method of multipliers algorithm to address the L_1 optimization problem.

[1] M. Arie-Nachimson, S. Z. Kovalsky, I. Kemelmacher-Shlizerman, A. Singer, and R. Basri. Global motion estimation from point matches. In *Proc. 3DPVT*, 2012.

[2] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[3] N. Jiang, Z. Cui, and P. Tan. A global linear method for camera pose registration. In *Proc. ICCV*, 2013.

[4] K. Wilson and N. Snavely. Robust global translations with 1dsfm. In *Proc. ECCV (3)*, pages 61–75, 2014.