Direct Structure Estimation for 3D Reconstruction

Nianjuan Jiang†, Wen-Yan Lin†, Minh N. Do‡, Jiangbo Lu†
†Advanced Digital Sciences Center, Singapore. ‡University of Illinois at Urbana-Champaign.

In this work we show that when combined with single/multiple homography estimation, the general Euclidean rigidity constraint provides a simple formulation for scene structure recovery without explicit camera pose computation. This direct structure estimation (DSE) opens a new way to design a SFM system that reverses the order of structure and motion estimation.

Almost all modern SFM systems start with relative pose estimation from feature correspondences (e.g. SIFT[5]) between two [2, 7] or three views [8, 9]. Reliable and accurate relative pose estimation is critical for a robust SFM system. However, to compute relative poses reliably is a non-trivial task. Most techniques suffer from instability caused by planar scenes [7], which is commonly seen in man-made environments. As a result, a separate process for detecting a dominant homography is often adopted in SFM systems.

Besides this well-known limitation of state-of-the-art SFM systems, there is also a technique 'void' in the general methodology of SFM as pointed out by Li [4]. In almost all traditional SFM methods, camera motion estimation always comes first, then followed by 3D structure computation.

While appreciating the rationales behind the traditional SFM schemes, such as theoretical elegance and practical effectiveness, we are interested in the feasibility and advantage of a structure-first approach for practical SFM systems. In fact, we observed that, with known intrinsic camera parameters, the ratio of the depths of a 3D point in two different views can be directly inferred from a homography relating the two image points (see Fig. 1).

We find the proposed approach works particularly well for sideway motion regardless of the number of available planar structures. This is actually a desired property in practice, since sideway motion is good for structure computation and is prevailing in data capturing for 3D reconstruction.

We use three views as the basic building block for DSE. The relative poses computed from the scene structures are readily integrated into existing SFM systems such as [3, 6]. If a pair of corresponding calibrated points $\mathbf{p} = (x, y, 1)^T$ and $\mathbf{p}' = (x', y', 1)^T$ in images I and I' are related by a homography $\mathbf{H}$, we have the following equation

$$\lambda \mathbf{p}' = \mathbf{H}\mathbf{p}, \tag{1}$$

where λ is a scalar. $\mathbf{H}$ is scaled such that $\mathbf{H} = \mathbf{R} + \frac{\mathbf{t}\mathbf{n}^T}{d_\pi}$.

Proposition: Let d and d' denote the depths of a 3D point $\mathbf{X}$ in view I and I', with projected 2D points $\mathbf{p}$ and $\mathbf{p}'$, respectively. Then we have the equality $\lambda = \frac{d'}{d}$. The proof is given in the paper.

We show that the homography induced depth ratio together with the Euclidean rigidity constraint lead to a simple formulation for solving the relative depths of 3D point pairs. According to the Euclidean rigidity constraint, the distance between the 3D points $\mathbf{X}_i$ and $\mathbf{X}_j$ does not change under any rigid body transformation, i.e. $\|d'_i\mathbf{p}'_i - d'_j\mathbf{p}'_j\| = \|d_i\mathbf{p}_i - d_j\mathbf{p}_j\|$.

Given the depth ratio $\lambda_i = \frac{d'_i}{d_i}$ and $\lambda_j = \frac{d'_j}{d_j}$ obtained from the respective homography relating each pair of the corresponding points, we can obtain

$$\|\lambda_i \frac{d_i}{d_j}\mathbf{p}'_i - \lambda_j \mathbf{p}'_j\| = \|\frac{d_i}{d_j}\mathbf{p}_i - \mathbf{p}_j\|. \tag{2}$$

Let $\alpha = \frac{d_i}{d_j}$, we arrive at the following quadratic equation about α,

$$\begin{aligned} A\alpha^2 + B\alpha + C &= 0, \text{where} \\ A &= \|\lambda_i \mathbf{p}'_i\|^2 - \|\mathbf{p}_i\|^2, \\ B &= -2(\lambda_i\lambda_j \mathbf{p}'^T_i \mathbf{p}'_j - \mathbf{p}_i^T \mathbf{p}_j), \\ C &= \|\lambda_j \mathbf{p}'_j\|^2 - \|\mathbf{p}_j\|^2. \end{aligned} \tag{3}$$

Given a third view I'', we directly solve the following minimization problem to obtain the optimal solution,

$$\alpha_{ij} = \arg\min_\alpha \left|A_1\alpha^2 + B_1\alpha + C_1\right| + \left|A_2\alpha^2 + B_2\alpha + C_2\right|. \tag{4}$$

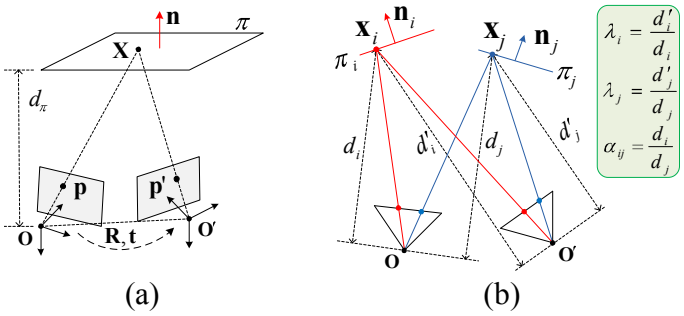

Figure 1: The proposed DSE utilizes the homography induced depth ratio and Euclidean rigidity constraint to estimate the structure directly without camera pose recovery. (a) Geometric interpretation of homography decomposition. (b) Homography induced depth ratios λ_i and λ_j together with the rigidity constraint give the estimate for α_{ij}.

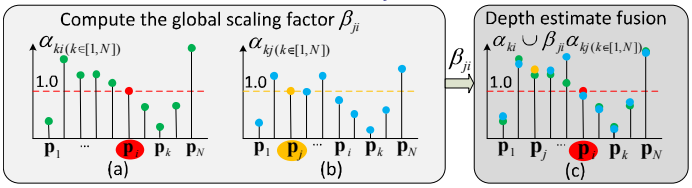

Figure 2: Structure estimation from two sets of relative depths (best viewed in color). (a) Relative depths α_{ki} computed using point $\mathbf{p}_i$ as reference. (b) Relative depths α_{kj} computed using point $\mathbf{p}_j$ as reference. (c) The final structure is computed as the average of the scaled relative depths.

Collectively, for each point $\mathbf{p}_i \in S$ with its depth fixed as $d_i = 1$, the depths of all the points in the same view are given by $\alpha_{ki} = \frac{d_k}{d_i}$. If there is zero noise in the data, we shall have

$$\{\alpha_{ki}\} = \beta_{ji}\{\alpha_{kj}\}, \forall k \in [1, N], \tag{5}$$

meaning that each set of depths only differs by a global scaling factor (see Fig. 2). We compute the average scaling factor for each set of depths using RANSAC [1]. The average depth for each point $\mathbf{p}_i$ is computed similarly after applying the scaling factor to each set of depth estimation (Fig. 2(c)).

The proposed techniques for multiple homography estimation and to integrate DSE to a general SFM system are presented in the paper.

[1] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[2] R. Hartley. In defense of the eight-point algorithm. *IEEE Trans. PAMI*, 19(6):580–593, 1997.

[3] N. Jiang, Z. Cui, and P. Tan. A global linear method for camera pose registration. In *Proc. ICCV*, pages 481–488, 2013.

[4] H. Li. Multi-view structure computation without explicitly estimating motion. In *Proc. CVPR*, pages 2777–2784, 2010.

[5] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[6] P. Moulon, P. Monasse, and R. Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *Proc. ICCV*, pages 3248–3255, 2013.

[7] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Trans. PAMI*, 26(6):756–770, 2004.

[8] D. Nistér and F. Schaffalitzky. Four points in two or three calibrated views: Theory and practice. *IJCV*, 67(2):211–231, 2006.

[9] L. Quan. Invariants of six points and projective reconstruction from three uncalibrated images. *IEEE Trans. PAMI*, 17(1):34–46, 1995.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.