Direct Structure Estimation for 3D Reconstruction

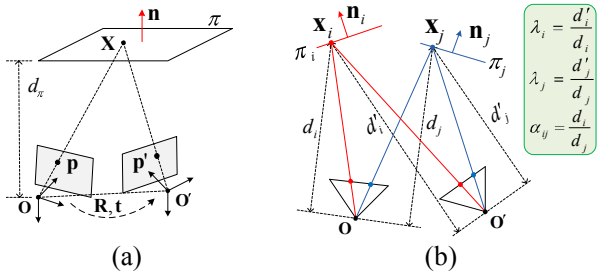
Nianjuan Jiang[†*], Wen-Yan Lin[†], Minh N. Do[‡], Jiangbo Lu[†*]
[†]Advanced Digital Sciences Center, Singapore
[‡]University of Illinois at Urbana-Champaign

Abstract

Most conventional structure-from-motion (SFM) techniques require camera pose estimation before computing any scene structure. 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. We show that this alternative approach works well for recovering scene structure and camera poses from sideway motion given planar or general man-made scenes.

1. Introduction

Structure from motion (SFM) is a classical problem in computer vision and has been studied actively for decades. In recent years, driven by the increasing demands of industrial applications such as navigation, augmented reality, robotics and film/game production, significant progresses have been made that advance the SFM techniques in terms of the system reliability and scalability [21, 24]. Almost all modern SFM systems start with relative pose estimation from feature correspondences (e.g. SIFT[15]) between two [8, 19] or three views [20, 22]. These relative poses will be merged into a global coordinate system afterwards [24, 13, 17, 4, 10]. The scene structure is then computed and refined together with all camera parameters, e.g. by bundle adjustment (BA) [29]. Therefore, 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 [19], 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. On the other hand, planar structure by itself actually gives a strong geometric constraint and can be utilized for better reconstruction quality [3, 11, 32].



(a) (b)

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}.

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 [14]. In almost all traditional SFM methods, camera motion estimation always comes first, then followed by 3D structure computation[1]. The work by Li [14] is among the earliest to propose an actual implementation that bypasses the motion estimation.

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(b) and Section 3.1). Furthermore, the *Euclidean rigidity constraint* implies that the Euclidean distance between two 3D points is invariant under a rigid body transformation. When combined with the aforementioned homography induced depth ratio across different views, we can derive a simple equation from the rigidity constraint to solve for the relative depths of two 3D points uniquely given at least three different views, i.e. the

*Corresponding authors: Nianjuan Jiang (nianjuan.jiang@adsc.com.sg), Jiangbo Lu (jiangbo.lu@adsc.com.sg)

[1]Tomasi-Kanade factorization is an exception which recovers structure and motion simultaneously.

scene structure can be determined directly up to a common scale from three calibrated images (see Fig. 1(b) and Section 3.2). Note that although our method involves homography estimation, we do not require the points to share the homography, nor do we estimate any camera parameters from the homography matrix. For easy reference, we term this approach as *direct structure estimation* (DSE), and will use it hereinafter.

To evaluate the potential of DSE for a practical SFM system, we further recover the camera parameters from the estimated structures. Specifically, we compute the scene structure in the camera view of each image, and obtain their relative poses by a simple 3D rigid body transformation [2, 5]. The estimation can be further refined using non-linear optimization, e.g. bundle adjustment (BA) [29]. 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.

2. Related work

Our work can be considered as one example of 'structure-first' SFM techniques. As compared to the rich body of SFM literature, this is a relatively 'void' space. Early works have speculated the feasibility for obtaining general 3D structure without explicitly computing camera motions (e.g. [7, 31]). Li [14] has proposed for the first time an actual implementation for such a scheme based on a graph rigidity theory, where a subset of the inter-point Euclidean distances are computed before embedding the actual coordinates of the 3D points. However, to extend such a scheme to a robust and scalable SFM system is not obvious. Inter-point distance has been used in early vision works to derive multi-view invariants (e.g. [7, 30]). Our method also utilizes the invariance of the inter-point distance under rigid body transformations to derive the constraints. Tardif *et al.* [26] used the factorization framework and proposed a structure basis constraint that can recover scene structure first mainly for affine cameras. Aliaga *et al.* [1] proposed a structure estimation scheme by eliminating motion parameters from the SFM formulation, but it requires initialization for the resulting nonlinear bundle adjustment problem.

It is well known that prior knowledge about the scene planes can greatly facilitate the 3D reconstruction problem. Plane-based camera self-calibration and 3D reconstruction from uncalibrated views have been well studied in the literature [28, 11, 3]. Zhou *et al.* [32] proposed a fully automatic SFM system based on dominant planes detected in the scene from an uncalibrated video sequence. While these works deal with uncalibrated images and aim to recover the camera motion and scene structure simultaneously from mutliple views, we show the feasibility to directly estimate scene structure from image correspondences related by homographies and its readiness as a component for a general scalable SFM system.

Our DSE method involves robust multiple homography detection, which is a challenging and active research topic [27, 9]. The objective of the classic problem is to cluster the image points such that they form a minimum number of co-planar regions and each region accurately covers as many points as possible. However, our objective is slightly different from the classic problem statement. In fact, we are not concerned whether the number of homographies detected is optimal, and the points can have multiple homography assignments simultaneously. This relaxation makes our problem much easier and we propose a simple method to achieve our goal.

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 [10, 18]. In particular, we use the open source code provided by Jiang *et al.* [10] to register the cameras globally and apply BA to obtain the final reconstruction.

3. Direct structure estimation

In the following, we first introduce the structure constraint induced by homography for calibrated cameras. This constraint gives us the knowledge of the ratio between the depths of a 3D point seen from two different views. Then, we shall use this depth ratio to derive the equation for solving the relative depths of two 3D points observed in the same view. In general, there are two valid solutions to the equation we derived. Hence, we propose to use a third view to resolve this ambiguity and produce a unique solution for every pair of 3D points. We also design a robust scheme to harvest the scene structure from all such pairwise estimations, which are usually contaminated by noise and error.

3.1. Homography induced structure constraints

We begin with the formal proof of the structure constraint induced by homography.

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.

Suppose $\mathbf{H}$ is scaled[2] such that $\mathbf{H} = \mathbf{R} + \frac{\mathbf{t}\mathbf{n}^T}{d_\pi}$, where $\mathbf{R}$ and $\mathbf{t}$ denote the camera rotation and translation between the two views, and π is the plane defined with $(\mathbf{n}, d_\pi)$ in the camera coordinate system of view I (see Fig. 1). Here, $\mathbf{n}$ denotes the normal of the plane and d_π denotes the distance of the plane π from the camera center of view I.

[2]The scaling factor is given by the second largest singular value of $\mathbf{H}$, see [16].

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 $\lambda = \frac{d'}{d}$.

Proof. Let $\mathbf{X}$ denote a 3D point on plane π, and satisfying the plane equation $\mathbf{X}^T\mathbf{n} - d_\pi = 0$. The camera projection matrices of view I and I' are given by $[\mathbf{I}\ \mathbf{0}]$ and $[\mathbf{R}\ \mathbf{t}]$, respectively. Given the depth d (d') of $\mathbf{X}$ in view I (I'), we have

$$d\mathbf{p} = \mathbf{X}, \tag{2}$$

$$d'\mathbf{p}' = \mathbf{R}\mathbf{X} + \mathbf{t}. \tag{3}$$

Substitute Eq. (2) to the plane equation and Eq. (3):

$$\frac{1}{d} = \frac{\mathbf{n}^T\mathbf{p}}{d_\pi}, \tag{4}$$

$$\frac{d'}{d}\mathbf{p}' = \mathbf{R}\mathbf{p} + \frac{\mathbf{t}}{d}. \tag{5}$$

Combine Eq. (4) and Eq. (5), we have

$$\frac{d'}{d}\mathbf{p}' = (\mathbf{R} + \frac{\mathbf{t}\mathbf{n}^T}{d_\pi})\mathbf{p} = \mathbf{H}\mathbf{p}, \Rightarrow \lambda = \frac{d'}{d}. \tag{6}$$

□

3.2. Relative depth recovery

In the following, 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.

Given two pairs of corresponding points $(\mathbf{p}_i, \mathbf{p}'_i)$ and $(\mathbf{p}_j, \mathbf{p}'_j)$, we denote their respective depths in view I and I' as (d_i, d'_i) and (d_j, d'_j). 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, with simple manipulation, we can obtain the following equation:

$$\|\lambda_i\frac{d_i}{d_j}\mathbf{p}'_i - \lambda_j\mathbf{p}'_j\| = \|\frac{d_i}{d_j}\mathbf{p}_i - \mathbf{p}_j\|. \tag{7}$$

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

$$A\alpha^2 + B\alpha + C = 0, \text{where} \tag{8}$$

$$A = \|\lambda_i\mathbf{p}'_i\|^2 - \|\mathbf{p}_i\|^2,$$

$$B = -2(\lambda_i\lambda_j\mathbf{p}'^T_i\mathbf{p}'_j - \mathbf{p}_i^T\mathbf{p}_j),$$

$$C = \|\lambda_j\mathbf{p}'_j\|^2 - \|\mathbf{p}_j\|^2.$$

We can easily solve the above equation and obtain up to two valid solutions for α. Equivalently, we obtain the relative depths of the corresponding 3D points $\mathbf{X}_i$ and $\mathbf{X}_j$ in view I.

Given a third view I'', there are up to two additional solutions for α, and we can select the one that satisfies both equations and is positive. In fact, we directly solve the following minimization problem to obtain the optimal solution[3],

$$\alpha_{ij} = \arg\min_\alpha \left|A_1\alpha^2 + B_1\alpha + C_1\right| + \left|A_2\alpha^2 + B_2\alpha + C_2\right|. \tag{9}$$

We use α_{ij} to denote the estimated relative depth ratio between $\mathbf{X}_i$ and $\mathbf{X}_j$ (Fig. 1 (b)). Here, (A_1, B_1, C_1) and (A_2, B_2, C_2) are coefficients computed from the view pair (I, I') and (I, I''), respectively.

In fact, Eq. (9) minimizes the average difference between the Euclidean distance between X_i and X_j measured in camera view I, I' and I'', respectively.

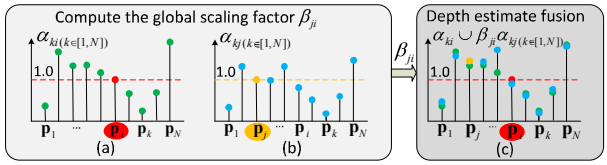


Figure 2. Structure estimation from two sets of relative depths (best viewed in color). (a) Relative depths α_{ki} computed using point $\mathbf{p}_i$ as reference. (b) Relative depths α_{kj} computed using point $\mathbf{p}_j$ as reference. (c) The final structure is computed as the average of the scaled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

3.3. Structure estimation

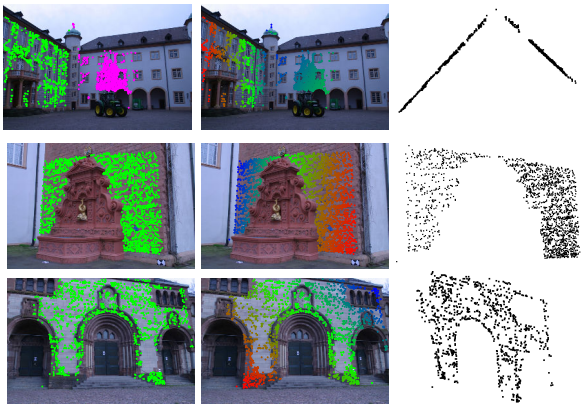
So far we have shown how to obtain the relative depths of two 3D points given their correspondences and associated homographies across three views. Ideally, one can fix the depth of an arbitrary point, and compute the rest to obtain the scene structure up to a global scale. In reality, the results will obviously be biased by the chosen reference point. Since the computation for α is simple and can be easily parallelized for different 3D point pairs, we do this exhaustively for all pairwise combination of 3D points that find correspondences across three images.

Now we denote the set of image points in view I that has correspondences in the other two images as $S = \{\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_N\}$, and N is the total number of such points. 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

$$\{\alpha_{ki}\} = \beta_{ji}\{\alpha_{kj}\}, \forall k \in [1, N], \tag{10}$$

meaning that each set of depths only differs by a global scaling factor (see Fig. 2). In the presence of noise, each pair of $(\alpha_{ki}, \alpha_{kj})$ will give a different estimate for β_{ji}. Therefore, we compute the average scaling factor for each set of depths using RANSAC [6] (the threshold is set as 1% of the expected value of β_{ji}). 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)).

[3]The minimization always gives a real solution while the original equation may have no real solutions.



(a) (b) (c)

Figure 3. Multiple homography detection and structure computation. From top to bottom we show the results for the dataset '*castle-P30*', '*fountain-P11* and '*Herz-Jesu-P25*' [25], respectively. (a) The representative co-planar point clusters. Points belonging to the same homography are marked in the same color. (b) The recovered depths colored according to their relative values. Red means near and blue means far. (c) Scene structure viewed in 3D.

In practice, we also compute α'_{ij} for views I', and propagate the values to view I using the point depth ratio across views given by homographies as $\tilde{\alpha}_{ij} = \alpha'_{ij}\frac{\lambda_j}{\lambda_i}$. The same computation is repeated for view I''. We take the average of all three $\tilde{\alpha}_{ij}$ as the final estimate for each α_{ij}. We discard the estimate for α_{ij} if the standard deviation of $\tilde{\alpha}_{ij}$ is large (e.g. 1% of the expected value) for better robustness. Figure 3(b) and (c) show examples of the typical structures we recover. Without estimating any camera motion parameters, we nicely recover the dominant planar structures in the scene. The orthogonal relationship between the two wall facades in '*castle-P30*' is well preserved.

3.4. Multiple homography estimation

Before applying DSE, we need to detect the presence of homographies and compute each homography transformation from image correspondences. This is a typical multi-model fitting problem, and there are sophisticated algorithms proposed for this task, e.g. [27, 9]. However, our requirements are slightly different and relatively relaxed as compared to the classic problem statement. First, we do not care whether the number of homographies discovered is optimal so long as each point cluster truely conforms to a homography. Second, each image point can be assigned to different homographies. This is in fact desired in our case, since each homography fitting will give an estimate to the depth ratio of a point across two views. We also like to point out that generally local planes exist everywhere and a homography is also a good approximation to many geometrically non-planar structures if fitted locally. Therefore, we adopt a simple 'fit-and-grow' approach by sweeping through evenly spaced image regions for homography detection. Recent work [23] also used a similar strategy to generate plane hypotheses for stereo matching.

In particular, we first divide the image plane of the reference view into overlapping cells (e.g. 50% overlap) with size $L \times L$[4]. Then we apply RANSAC to generate a homography hypothesis within each cell and use all the correspondences found in the second image for inlier/outlier testing (we set the threshold as 2 pixels for image resolution of 1600×1200). We accept a homography hypothesis as valid only if the number of inliers exceeds 10. For each successful hypothesis, we repeat hompography fitting using all inliers and perform inlier/outlier testing on all correspondences until no more new inliers are found. We optimize the final homography by minimizing the Sampson's error. By performing the 'growing' step, non-local inliers can also be aggregated. This is useful for spatially unconnected but geometrically co-planar surfaces or bulding facades. We show example results of the proposed technique in Fig. 3 (a).

In our case, an image point can participate in more than one homography fitting. Typically, the number of homopgrahies fitted for each image point ranges from 0 to 6 in the examples we tested. The difference between the individual depth ratio estimate and the averaged value is usually less than 1%. Since homographies estimated with more points are often more accurate and stable, we weight each λ by the number of inliers used for its homography estimation, and take the weighted average for our computation.

4. Integration with SFM systems

An immediate application of DSE is to serve as a building block for a general SFM system. Given a collection of images, we can apply DSE to every view triplet with sufficient overlap (e.g. by considering the number of common correspondences found between them). The relative pose between views within a triplet can be obtained by computing a 3D rigid transformation between the scene structure recovered for each view. These relative poses are readily fed into SFM systems such as [18, 10]. We will describe each step in detail in the following.

Once we obtain the scene structure from homography detection and DSE, we can recover camera poses as a side product by 3D rigid body transformation using SVD [2, 5]. We did not use the standard camera absolute pose algorithm given 2D-3D correspondences (e.g. EPnP [12]) because the PnP algorithm also suffers from instability in the presence of a single planar structure. In fact, we find the 3D rigid body transformation gives comparable results on camera pose estimation in general. In practice, for the best results, we can use these initial camera poses to triangulate the remaining image correspondences that are not recovered in

[4]The cell size L is given as $1/10$ of the larger dimension of the image.

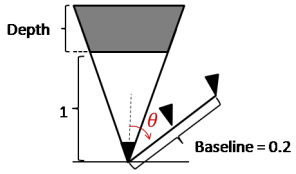


Figure 4. Camera and scene setup for synthetic experiments.

the DSE step and refine the camera poses by BA.

Given more than three images, we first recover the relative camera poses from the computed scene structure for each view triplet. We then feed these relative poses to the algorithm proposed by Jiang *et al.* [10] and produce the 3D reconstruction for multiple images.

5. Experiments

We evaluate DSE with both synthetic data and real data to fully understand its behavior and potential in SFM applications. We compare DSE with three representative calibrated relative pose algorithms on scenes with synthetic planar structure(s). Namely, we choose the direct homography decomposition algorithm (HD) [16], the 5-point algorithm (2V5P) for epipolar geometry [19] and the four-point algorithm (3V4P) [20] for trifocal tensor. For simplicity, we only test HD for the case of a single planar structure.

For synthetic experiments, we follow the conventional set up as described in previous literature, e.g. [19, 20]. As shown in Fig. 4, the first camera is oriented to align with the world coordinate system. The second camera is placed at 0.1 units away from the first camera, and the third camera is sitting in the middle of the baseline between the previous two cameras. The direction of camera translation is controlled by the angle θ, e.g. $\theta = 90°$ corresponds to sideway motion. The second and the third camera is rotated such that its optical axis passes through the centroid of the imaged points, with the x-axis remaining parallel to the x-z plane and the y-axis pointing to the same half-space as the world y-axis. The horizontal field of view of the camera is $45°$ and the image resolution is 352×288 in pixels. We perturb the image coordinates by zero-mean Gaussion noise with different standard deviations.

The scene points are generated within the view frustum of the first camera with a minimum depth of 1 unit and scene depth of 0.5 units. In the case of planar scene, the plane is generated such that it passes through the center of the scene frustum and its normal deviates from the z-axis by an angle of 0 to 30 degrees. In the case of multiple planes, the plane orientations are generated similarly to the single plane case. The location of each plane is determined by assigning an arbitrary point in the scene frustum to it. Randomly sampled scene points are arbitrarily projected to the visible parts of these planes. We generate in total five different planes for the test. Note that it is not easy to simulate realistic piecewise planar scenes without introducing bias. As we are only interested in the algorithm behavior of DSE, perfect clustering of the points is given for homography fitting in all cases. All the parameters involved in our computation are still estimated from the given noisy data.

5.1. Accuracy and stability of DSE

We test the accuracy and stability of DSE over different camera translation directions under varying noise levels. Among all the translation directions, forward and sideway motion are two special motions that are often encountered during data capturing. However, as a rule of thumb, for the purpose of reconstruction, forward motion is typically not recommended since the triangulation of scene points can be extremely sensitive to image noise and small camera motion errors. This disadvantage to structure recovery was also observed in our experiments, however, with an alternative explanation given later. Therefore, the performance of the algorithm on sideway motion is of more importance for the reconstruction purpose in practice. We use the publically available source code for 2V5P, and our own implementation of the 3V4P and HD for the comparison. Since HD generally gives up to two valid decompositions [16], we disambiguate the results by finding the common plane normal recovered by the two homographies between the reference view and the other two views.

The comparisons of camera pose estimation accuracy are given in Fig. 5 and Fig. 6. Each data point in these figures is computed over 100 trials. The relative rotation error and translation error in degrees are shown in the first two columns respectively. These errors are computed over 'inlier' pose estimates where the relative rotation error is smaller than $3°$. The ratio of the camera pose estimates with gross error is given in the third column. By doing this, we have a better understanding of how good an algorithm is at obtaining the correct solution and its actual numerical stability to image noise.

We found that both 2V5P and 3V4P are very unstable with planar scenes in the presence of image noise (Fig. 5(c)), especially in the case of sideway motion (Fig. 5(a)). On the other hand, DSE constantly produces the best results for sideway motion regardless of the number of available homographies (Fig. 5(a) and Fig. 6(a)). HD shows good performance for sideway motion when the image noise is low, but correct plane normal selection becomes more difficult as image noise increases. The performance of DSE on forward motion, however, is not on-par with its performance on sideway motion. It seems the algorithm has difficulty in solving for the correct structure stably when the image noise becomes large. We observe the same behavior with HD. With a more careful inspection, we find that the displacement of the image points can often be well explained by a homography induced by a frontal parallel

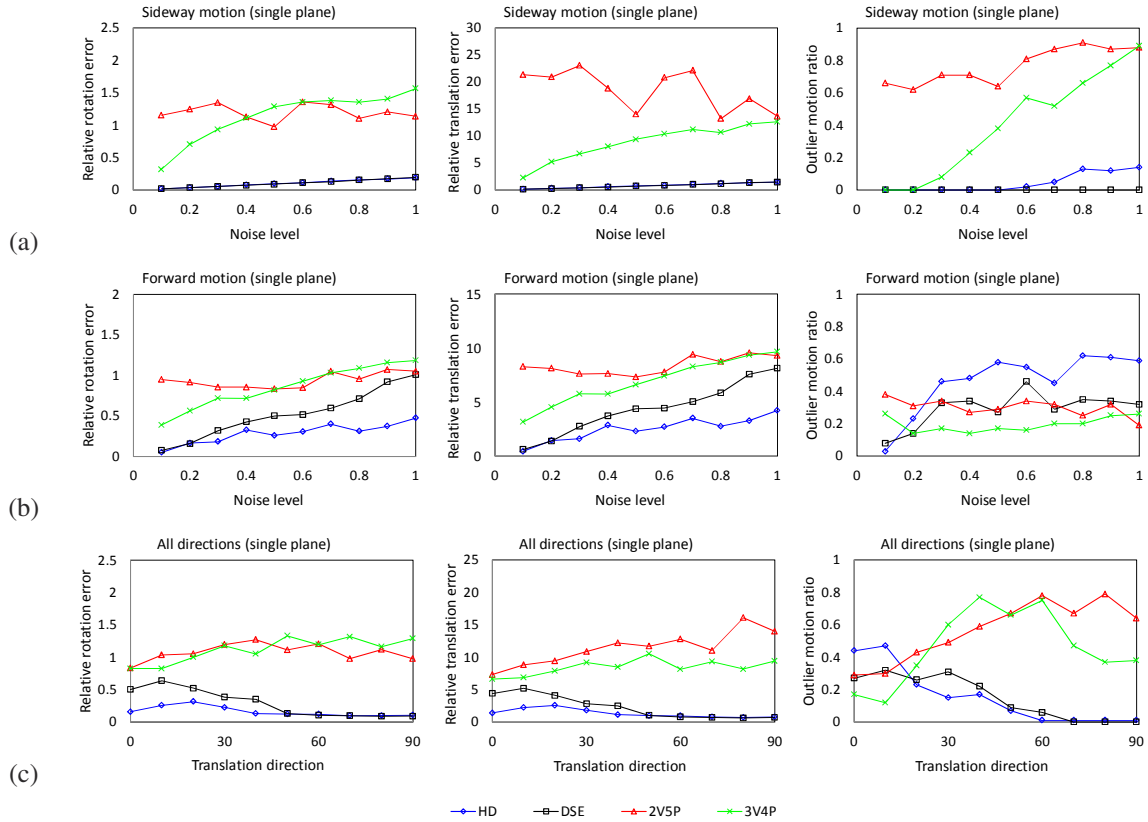


Figure 5. Relative rotation error and translation error for scene with a single plane over different translation directions. (a) Sideway motion. (b) Forward motion. (c) All directions with a noise level of 0.5 pixels.

plane when the camera undergoes forward motion with a relatively small baseline. All the outlier motions produced by DSE are resulting from this particular 'false structure'. We give such an example in Fig. 7(a).

When the scene contains multiple planes, 2V5P generates good results for all kinds of motions and constantly outperforms 3V4P. DSE gives the best results for sideway motion, yet it still suffers from structure confusion for forward motion. The reason is similar to the single plane case.

The current DSE computes the optimal relative depth α independently for each point without considering the consistency with other points. We believe this global consistency is the key to resolve the structure confusion in forward motion. Therefore, an interesting future direction is to consider the consistency of the relative depths among all the points and choose the configuration that minimizes the reprojection error over all the image observations. Nevertheless, we can see from Fig. 5(c) and Fig. 6(c) that the comfortable operating zone for the current DSE ranges from $50°$ to $90°$ *regardless* of the type of scene structure. We consider this as a complementary algorithm behavior as compared to the standard relative pose algorithms such as 2V5P.

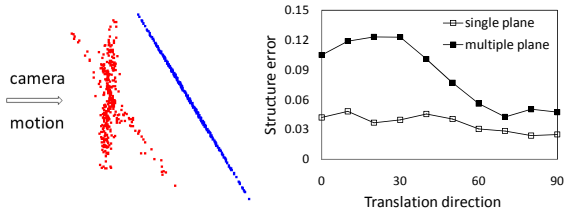


Figure 7. (a) An example of the recovered frontal parallel 'false structure' (colored in red) given by DSE when camera undergoes forward motion. The true structure is colored in blue. (b) The structure estimation error of DSE for different types of camera motion (image noise level is 0.5 pixels).

5.2. 3D reconstruction

When integrating DSE into a SFM system such as [10], we need to remove false triplet reconstructions. Here, triplet verification obviously does not work since the false structure and camera poses usually consistitute an ambiguous solution. Instead, we perform pairwise verification. For each view pair, we can compare the relative pose estimates between them obtained from different view triplets to identify outliers. We simply consider a relative pose (and hence the triplet it comes from) as an outlier if its minimum rotation

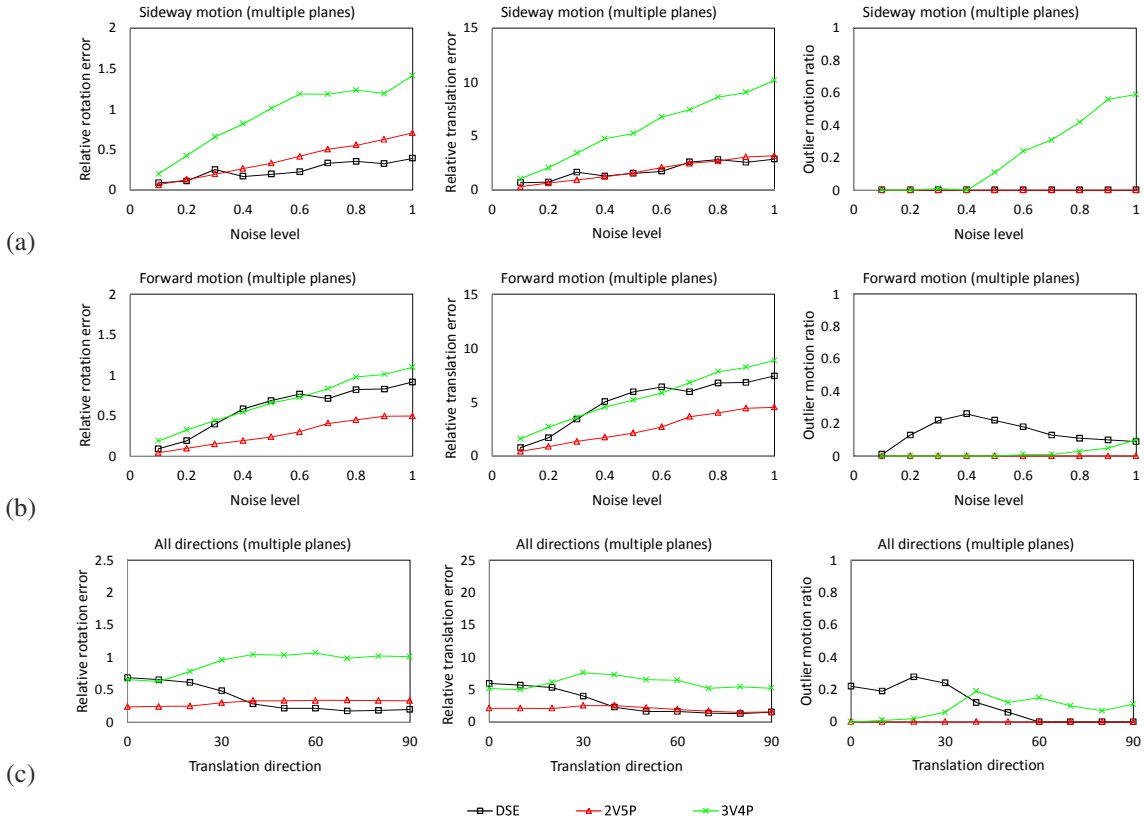
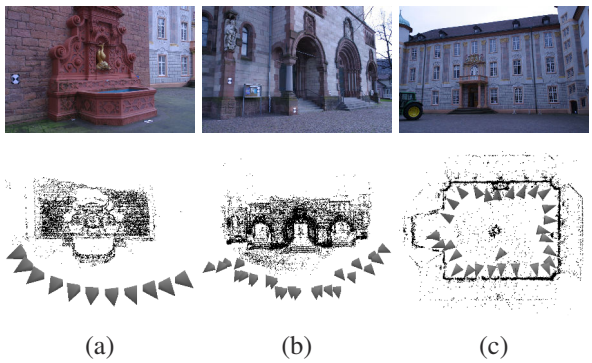


Figure 6. Relative rotation error and translation error for scene with multiple planes over different translation directions. (a) Sideway motion. (b) Forward motion. (c) All directions with a noise level of 0.5 pixels.

difference from at least two other solutions is greater than 3°. We test the DSE-based SFM system with six real image sequences. We use the benchmark dataset '*fountain-P11*', '*Herz-Jesu-P25*' and '*castle-P30*'[25] to provide a quantitative evaluation. The feature correspondences are computed using SIFT [15]. The view triplets used for the computation are generated by first connecting each image with three other images with most correspondences, and then collecting all the triplets formed by these view pairs. The results are reported in Table 1. We test our algorithm with both ground truth calibration (GT) and calibration read from the Exif tags (Exif). Here, $R3_{err}$ and $t3_{err}$ denote the average relative rotation error and translation error in degrees within view triplets, respectively. The average error in absolute rotation (in degrees) and camera position (in cm) before the final BA are given by R_{err} and c_{err}, respectively. The absolute camera position error after the final BA is given by c_{err} (BA). The reconstruction obtained with DSE(SVD) in Table 1 using Exif calibration after the final BA is visualized in Fig. 8. For reference purpose, we also report the results obtained using [10].

Interestingly, we produce comparable results to [10] on these benchmark datasets even without applying three-view BA. In particular, we obtain a better camera pose initialization for '*castle-P30*'. This dataset contains images dominated by planar building facade and the 5-point algorithm produces large errors for relative pose estimation between those images pairs. In general, three-view BA should be applied to the initial camera poses obtained from DSE(SVD) to ensure the best initialization for the final BA.



(a) (b) (c)

Figure 8. Reconstruction results for benchmark datasets [25] using Exif information (without three-view BA). (a) *fountain-P11*. (b) *Herz-Jesu-P25*. (c) *castle-P30*.

We compare two more 3D reconstructions obtained using DSE-based SFM (with three-view BA) and the original 5-point based method in [10] visually in Fig. 9. The '*Street*'

fountain-P11	#Images	#Triplets	GT					Exif				
			$R3_{err}$	$t3_{err}$	R_{err}	c_{err}	c_{err} (BA)	$R3_{err}$	$t3_{err}$	R_{err}	c_{err}	c_{err} (BA)
DSE(SVD)	11	23	0.205	0.154	0.25	1.7	0.28	0.53	0.56	0.74	4.4	**1.1**
DSE(BA_3)			**0.09**	**0.066**	**0.021**	**0.9**	0.28	**0.32**	**0.49**	**0.45**	7.2	**1.1**
LinearSFM[10]			0.13	0.23	0.07	24	**0.27**	0.35	**0.49**	0.48	**3.4**	**1.1**

Herz-Jesu-P25	#Images	#Triplets	GT					Exif				
			$R3_{err}$	$t3_{err}$	R_{err}	c_{err}	c_{err} (BA)	$R3_{err}$	$t3_{err}$	R_{err}	c_{err}	c_{err} (BA)
DSE(SVD)	25	120	0.1	0.31	0.07	4.6	**0.6**	0.27	0.71	0.5	6.9	5.6
DSE(BA_3)			**0.057**	**0.17**	**0.06**	**1.2**	**0.6**	**0.17**	**0.47**	**0.39**	**6.1**	5.6
LinearSFM[10]			0.14	0.49	0.13	3	**0.6**	0.27	0.71	0.44	8.8	**5.5**

castle-P30	#Images	#Triplets	GT					Exif				
			$R3_{err}$	$t3_{err}$	R_{err}	c_{err}	c_{err} (BA)	$R3_{err}$	$t3_{err}$	R_{err}	c_{err}	c_{err} (BA)
DSE(SVD)	30	108	0.35	1.21	0.91	**45**	**10**	0.56	2.48	1.71	**158**	**20**
DSE(BA_3)			**0.22**	**0.62**	**0.27**	104	**10**	**0.35**	**1.24**	**0.96**	162	**20**
LinearSFM[10]			0.41	1.4	0.7	75	**10**	0.56	1.75	2.28	206	22

Table 1. Quantitative evaluation with benchmark datasets. 'GT' stands for ground truth calibration and 'Exif' stands for Exif calibration. 'BA_3' means three-view bundle adjustment.

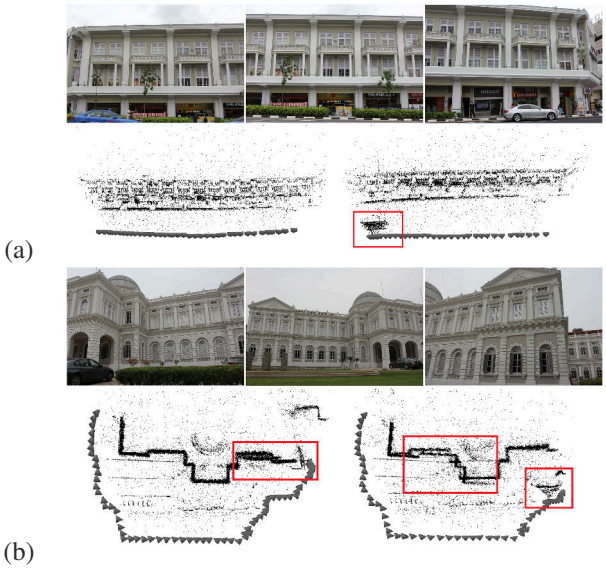


Figure 9. Reconstruction results for (a) *Street* and (b) *Building*. Below example images from each data sequence, we show reconstructions obtained by our DSE-based method on the left and the ones obtained by [10] on the right.

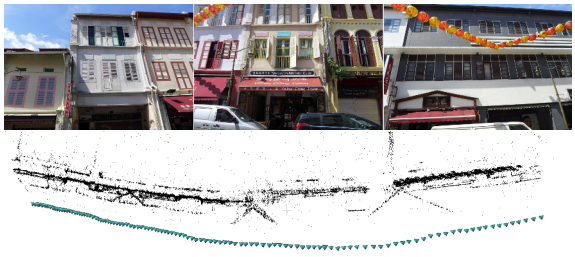


Figure 10. Example input images and the reconstruction for *Shophouse*.

sequence has 38 images and we collect 105 triplets by connecting neighboring images according to the time stamps to avoid confusion caused by repetitive structures. The '*Building*' example has 67 images and we collect 193 triplets. We use the same set of triplets for both methods. We can clearly see from Fig. 9 that due to noisy relative pose estimation by the 5-point algorithm, [10] produced misaligned reconstructions. The DSE-based SFM gives much better results, though it also suffers from poor relative pose estimation on a few view triplets of the '*Building*' example. This is most likely due to the presence of near forward motion with small baselines. We also show a reconstruction of the '*Shophouse*' sequence containing 122 images obtained by our DSE-based SFM in Fig. 10.

Our current unoptimized Matlab implementation of DSE takes about 8 seconds on a 2.53Hz CPU for a typical image triplet of size 1600×1200 dominated by piecewise planar scenes. Since all the computation involved in DSE is lightweight, speed-up is trivial.

6. Conclusion and future work

In this work, we show that given three calibrated images and scenes with detectable planes, we can directly estimate the structure without computing any camera motion parameters. This interesting discovery leads to a SFM scheme that is built on the reversed order, i.e., compute the structure first and then followed by pose estimation. Experimental results demonstrated that this structure computation is especially well suited for sideway motion regardless of the type of scene structures. This complementary algorithm behavior as compared to conventional relative pose algorithms opens a new way to think about the design of a robust SFM system. We believe there are ample rooms for improvement of the DSE-based SFM scheme. For instance, one can improve its performance by considering solving the relative depth globally, and utilizing lines for homography detection and fitting when dealing with indoor environments. Last but not least, combining DSE and conventional relative pose estimation for maximum stablility and versatility is by itself an interesting topic for investigation.

Acknowledgement

This study is supported by the HCCS research grant at the ADSC from Singapore's Agency for Science, Technology and Research (A*STAR).

References

[1] D. G. Aliaga, J. Zhang, and M. Boutin. Simplifying the reconstruction of 3d models using parameter elimination. In *Proc. ICCV*, pages 1–8, 2007. 2

[2] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-d point sets. *IEEE Trans. PAMI*, (5):698–700, 1987. 2, 4

[3] A. Bartoli and P. Sturm. Constrained structure and motion from multiple uncalibrated views of a piecewise planar scene. *IJCV*, 52(1):45–64, 2003. 1, 2

[4] D. Crandall, A. Owens, N. Snavely, and D. Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. In *Proc. CVPR*, pages 3001–3008, 2011. 1

[5] D. W. Eggert, A. Lorusso, and R. B. Fisher. Estimating 3-d rigid body transformations: a comparison of four major algorithms. *Machine Vision and Applications*, 9(5-6):272–290, 1997. 2, 4

[6] 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. 3

[7] N. M. Grzywacz and E. C. Hildreth. Incremental rigidity scheme for recovering structure from motion: Position-based versus velocity-based formulations. *J. Opt. Soc. Am. A*, 4(3):503–518, 1987. 2

[8] R. Hartley. In defense of the eight-point algorithm. *IEEE Trans. PAMI*, 19(6):580–593, 1997. 1

[9] H. Isack and Y. Boykov. Energy-based geometric multi-model fitting. *IJCV*, 97(2):123–147, 2012. 2, 4

[10] N. Jiang, Z. Cui, and P. Tan. A global linear method for camera pose registration. In *Proc. ICCV*, pages 481–488, 2013. 1, 2, 4, 5, 6, 7, 8

[11] R. Kaucic, R. Hartley, and N. Dano. Plane-based projective reconstruction. In *Proc. ICCV*, volume 1, pages 420–427, 2001. 1, 2

[12] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o(n) solution to the pnp problem. *IJCV*, 81(2):155–166, 2009. 4

[13] M. Lhuillier and L. Quan. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE Trans. PAMI*, 27(3):418–433, 2005. 1

[14] H. Li. Multi-view structure computation without explicitly estimating motion. In *Proc. CVPR*, pages 2777–2784, 2010. 1, 2

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 1, 7

[16] Y. Ma, S. Soatto, J. Kosecká, and S. S. Sastry. An invitation to 3-d vision: From images to geometric models (interdisciplinary applied mathematics). 2005. 2, 5

[17] D. Martinec and T. Pajdla. Robust rotation and translation estimation in multiview reconstruction. In *Proc. CVPR*, pages 1–8, 2007. 1

[18] 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. 2, 4

[19] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Trans. PAMI*, 26(6):756–770, 2004. 1, 5

[20] D. Nistér and F. Schaffalitzky. Four points in two or three calibrated views: Theory and practice. *IJCV*, 67(2):211–231, 2006. 1, 5

[21] M. Pollefeys, D. Nistér, J.-M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S.-J. Kim, P. Merrell, et al. Detailed real-time urban 3d reconstruction from video. *IJCV*, 78(2-3):143–167, 2008. 1

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

[23] S. N. Sinha, D. Scharstein, and R. Szeliski. Efficient high-resolution stereo matching using local plane sweeps. In *Proc. CVPR*, pages 1582–1589, 2014. 4

[24] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *IJCV*, 80(2):189–210, 2008. 1

[25] C. Strecha, W. V. Hansen, L. V. Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Proc. CVPR*, pages 1–8, 2008. 4, 7

[26] J.-P. Tardif, A. Bartoli, M. Trudeau, N. Guilbert, and S. Roy. Algorithms for batch matrix factorization with application to structure-from-motion. In *Proc. CVPR*, pages 1–8, 2007. 2

[27] R. Toldo and A. Fusiello. Robust multiple structures estimation with j-linkage. In *Proc. ECCV*, pages 537–547. 2008. 2, 4

[28] B. Triggs. Autocalibration from planar scenes. In *Proc. ECCV*, pages 89–105. 1998. 2

[29] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustmenta modern synthesis. In *Vision algorithms: theory and practice*, pages 298–372. 2000. 1, 2

[30] S. Ullman. Maximizing rigidity: the incremental recovery of 3-d structure from rigid and rubbery motion. 1983. 2

[31] M. Werman and A. Shashua. The study of 3d-from-2d using elimination. In *Proc. ICCV*, pages 473–479, 1995. 2

[32] Z. Zhou, H. Jin, and Y. Ma. Robust plane-based structure from motion. In *Proc. CVPR*, pages 1482–1489, 2012. 1, 2